

**РОССИЙСКАЯ АКАДЕМИЯ НАУК  
СИБИРСКОЕ ОТДЕЛЕНИЕ**

Федеральное государственное бюджетное научное учреждение

Федеральный исследовательский центр

Институт цитологии и генетики

На правах рукописи

**Соколов Владимир Сергеевич**

**КОМПЬЮТЕРНОЕ ИССЛЕДОВАНИЕ КОНТЕКСТНЫХ  
ХАРАКТЕРИСТИК ОТКРЫТЫХ РАМОК СЧИТЫВАНИЯ, СВЯЗАННЫХ С  
ЭФФЕКТИВНОСТЬЮ ЭЛОНГАЦИИ ТРАНСЛЯЦИИ, У  
ОДНОКЛЕТОЧНЫХ ОРГАНИЗМОВ**

03.01.09. - математическая биология, биоинформатика

Диссертация на соискание ученой степени

кандидата биологических наук

Научный руководитель:

к. б. н., с. н. с.

Матушкин Юрий Георгиевич

Новосибирск – 2015

## ОГЛАВЛЕНИЕ

<b>СПИСОК СОКРАЩЕНИЙ .....</b>	<b>6</b>
<b>ВВЕДЕНИЕ .....</b>	<b>7</b>
<b>ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ .....</b>	<b>14</b>
<b>1.1. Механизм процесса трансляции на примере трансляции <i>Escherichia coli</i>.....</b>	<b>14</b>
<b>1.2. Контекстные особенности открытых рамок считывания, связанные с эффективностью элонгации трансляции .....</b>	<b>16</b>
1.2.1. Влияние кодонного состава открытых рамок считывания на эффективность элонгации трансляции .....	16
1.2.1.1. Неравномерность использования синонимичных кодонов в открытых рамках считывания.....	16
1.2.1.2. Молекулярные механизмы и математические модели, объясняющие неравномерность использования кодонов и ее корреляцию с пулом тРНК.....	18
1.2.1.3. Связь функции гена с адаптацией его кодонного состава к пулу тРНК .....	23
1.2.1.4. Медленные кодоны в 5'-районе кодирующей части генов. Модель взаимодействия рибосомы с тРНК.....	23
1.2.1.5. Связь между скоростью трансляции и точностью синтеза и фолдинга белка.....	26
1.2.1.6. Кодонный состав генов и тканеспецифичная экспрессия у многоклеточных организмов .....	28
1.2.2. Влияние вторичных структур в мРНК на эффективность элонгации трансляции.....	29
1.2.2.1. Вторичные структуры в 5'-НТР и районе старт-кодона трансляции.....	29
1.2.2.2. Вторичные структуры в кодирующей части мРНК.....	31
1.2.2.3. Современные экспериментальные методы определения вторичных структур в мРНК .....	34
<b>1.3. Биоинформатические методы оценки эффективности экспрессии генов .....</b>	<b>35</b>
1.3.1. Математические индексы .....	35
1.3.2. Программы для анализа контекстных характеристик нуклеотидных последовательностей .....	40
1.3.2.1. Программы для анализа кодонного состав нуклеотидных последовательностей....	40

1.3.2.2. Программы для анализа вторичных структур в нуклеотидных последовательностях .....	42
1.3.3. Изучение процесса трансляции при помощи математических стохастических моделей .....	43
<b>Заключение к обзору литературы.....</b>	<b>47</b>
<b>ГЛАВА 2. МЕТОДЫ И АЛГОРИТМЫ.....</b>	<b>48</b>
<b>2.1. Геномные последовательности.....</b>	<b>48</b>
<b>2.2. Индекс эффективности элонгации трансляции EEI.....</b>	<b>48</b>
2.2.1. Учет кодонного состава гена при расчете индекса EEI.....	50
2.2.2. Учет потенциальных вторичных структур в мРНК при расчете индекса EEI.....	51
2.2.2.1. Индекс локальной комплементарности LCI .....	52
2.2.2.2. Индекс локальной комплементарности индивидуального нуклеотида $LCI(i, j)$ .....	54
2.2.3. Пять типов EEI и определение типа, лучше всего оценивающего эффективность элонгации трансляции в исследуемом организме.....	55
<b>2.3. Программа EloE .....</b>	<b>56</b>
2.3.1. Общие сведения о программе EloE.....	56
2.3.2. Входные и выходные данные программы EloE.....	59
2.3.3. Алгоритм работы программы EloE .....	63
<b>2.4. Статистический анализ .....</b>	<b>64</b>
<b>ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ .....</b>	<b>65</b>
<b>3.1. Исследование геномов одноклеточных организмов при помощи программы EloE. Расчет индексов эффективности элонгации трансляции .....</b>	<b>65</b>
3.1.1. Исследование геномов бактерий.....	65
3.1.2. Исследование геномов архей.....	66
3.1.3. Исследование геномов одноклеточных эукариот.....	67
3.1.4. Особенности организмов, выявленные в результате анализа геномов одноклеточных организмов программой EloE .....	68
<b>3.2. Подробное исследование организмов, принадлежащих к роду <i>Mycoplasma</i> .....</b>	<b>69</b>
3.2.1. Распределение исследованных штаммов <i>Mycoplasma</i> по пяти типам индекса EEI .....	69

3.2.2. Анализ количества совершенных локальных инвертированных повторов в генах различных штаммов <i>Mycoplasma</i> .....	72
3.2.3. Филогенетический анализ исследуемых <i>Mycoplasma</i> .....	74
3.2.4. Анализ профилей LCI индексов индивидуальных нуклеотидов у <i>Mycoplasma</i> .....	76
3.2.5. Подробное исследование профилей LCI индивидуальных нуклеотидов у <i>Mycoplasma</i> .....	78
3.2.6. Исследование оперонной структуры генов различных штаммов <i>Mycoplasma</i> .....	81
3.2.7. Связь между GC-составом и эволюционной оптимизацией первичной структуры генов <i>Mycoplasma</i> для повышения эффективности элонгации трансляции .....	82
3.2.8. Использование программы UNAFold для предсказания вторичной структуры мРНК у <i>Mycoplasma</i> .....	84
<b>3.3. Исследование оптимизации первичной структуры генов архей в процессе эволюции</b>	<b>86</b>
3.3.1. Исследование влияния синонимичных замен в генах организма на значения индекса EEI .....	86
3.3.2. Анализ предковых последовательностей генов архей .....	91
3.3.3. Исследование зависимости между влиянием потенциальных вторичных структур в мРНК на эффективность трансляции у архей и температурой их среды обитания .....	94
3.3.4. Связь между GC-составом и эволюционной оптимизацией первичной структуры генов архей для повышения эффективности элонгации трансляции .....	97
<b>3.4. Исследование взаимосвязи между эффективностью элонгации трансляции генов дрожжей и плотностью их нуклеосомной упаковки в 5'-фланкирующем районе</b>	<b>98</b>
3.4.1. Корреляция между потенциалом формирования нуклеосом и EEI у <i>S. pombe</i> .....	98
3.4.2. Корреляция между потенциалом формирования нуклеосом и EEI у <i>S. cerevisiae</i> .....	100
3.4.3. Корреляция между EEI и экспериментальными данными по нуклеосомной упаковке у <i>S. cerevisiae</i> .....	103
<b>3.5. Исследование возможных причин различия видов корреляций между ПФН и EEI у <i>S. cerevisiae</i> и <i>S. pombe</i></b>	<b>104</b>
3.5.1. Исследование профилей нуклеосомного потенциала у <i>S. cerevisiae</i> и <i>S. pombe</i> .....	104
3.5.2. Распределение генов по длине у <i>S. cerevisiae</i> и <i>S. pombe</i> .....	105
3.5.3. Исследование связи между GC-составом генов и индексом EEI у <i>S. cerevisiae</i> и <i>S. pombe</i> .....	106
3.5.4. Сравнение кодонных составов генов <i>S. cerevisiae</i> и <i>S. pombe</i> .....	107
3.5.5. Сравнение эффективностей элонгации трансляции генов с одинаковыми идентификаторами у <i>S. cerevisiae</i> и <i>S. pombe</i> при помощи программы EloE .....	108



3.5.6. Сравнение динуклеотидных составов генов <i>S. cerevisiae</i> и <i>S. pombe</i> .....	110
3.5.7. Результаты исследования возможных причин различия видов корреляций между ПФН и EEI у <i>S. cerevisiae</i> и <i>S. pombe</i> .....	113
<b>ЗАКЛЮЧЕНИЕ .....</b>	<b>115</b>
<b>ВЫВОДЫ.....</b>	<b>117</b>
<b>СПИСОК ЛИТЕРАТУРЫ.....</b>	<b>118</b>
<b>ПРИЛОЖЕНИЕ.....</b>	<b>136</b>
<b>Таблица 1. Значения типов индекса EEI для 2582 организмов из домена Бактерии.....</b>	<b>136</b>
<b>Таблица 2. Значения типов индекса EEI для 165 организмов из домена Археи.....</b>	<b>147</b>
<b>Таблица 3. Значения типов индекса EEI для 73 особенных организмов со значением параметра <math>M &lt; 30</math>.....</b>	<b>149</b>
<b>Таблица 4. Типы индексов для 62 штаммов <i>Mycoplasma</i>.....</b>	<b>150</b>
<b>Таблица 5. Профили средних по всем генам значений LCI индивидуальных нуклеотидов 62 исследованных микоплазм.....</b>	<b>153</b>
<b>Таблица 6. Список проанализированных таксономических групп архей и виды, принадлежащие этим группам. ....</b>	<b>157</b>
<b>Таблица 7. Среднее значение оптимальной температуры среды обитания (<math>T_{opt}</math>) 135 видов архей. ....</b>	<b>160</b>

## СПИСОК СОКРАЩЕНИЙ

- Анти-ШД** – последовательность, комплементарная последовательности Шайна-Дальгарно;
- БД** – база данных;
- ГРБ** – гены рибосомных белков;
- ГТФ** – гуанозинтрифосфат;
- МПР** – модель потока рибосом;
- НИК** – неравномерность использования кодонов;
- нт** – нуклеотид;
- НТР** – не транслируемый район;
- ОРС** – открытая рамка считывания;
- ПТЦ** – пептидил-трансферазный центр;
- ПФН** – потенциал формирования нуклеосом;
- ШД** – последовательность Шайна-Дальгарно;
- САИ** – индекс адаптации кодонов (codon adaptation index);
- СБИ** – индекс смещения кодонов (codon bias index);
- CDS** – белок-кодирующая последовательность (coding sequence);
- ЕЕИ** – индекс эффективности элонгации (elongation efficiency index);
- ENc** – эффективное число кодонов (effective number of codons);
- fMET-тРНК** – формилметионин-тРНК;
- F<sub>оп</sub>** – частота оптимальных кодонов (frequency of optimum codons);
- LCI** – индекс локальной комплементарности (local complementary index);
- LCI(*i*, *j*)** – индекс локальной комплементарности индивидуального нуклеотида *j* в гене *i*;
- PARS** – параллельный анализ вторичной структуры РНК (parallel analysis of RNA structure);
- RFM** – модель потока рибосом (ribosomal flow model).

## ВВЕДЕНИЕ

### Актуальность темы исследования

Трансляция – это процесс синтеза белка из аминокислот на матрице информационной (матричной) РНК (мРНК), осуществляемый рибосомой. Это очень сложный, многостадийный процесс, в котором принимает участие огромное количество разнообразных молекул. Выделяют три основных стадии трансляции: инициацию, элонгацию и терминацию. Стадия инициация считается лимитирующим звеном трансляции [Kaczanowska and Rydén-Aulin, 2007]. Однако после инициации, элонгация является самой время- и энергозатратной. Время прохождения каждой из стадий вносит свой вклад в суммарное время трансляции. Соответственно, скорость синтеза белка – эффективность трансляции, зависит от эффективности каждой стадии.

На эффективность стадий трансляции оказывают влияние различные факторы. Например, для прокариот показана связь эффективности инициации с наличием в районе старта кодона трансляции определенной последовательности нуклеотидов, последовательности Шайна-Дальгарно (ШД) [Kaczanowska and Rydén-Aulin, 2007]. Другие исследования показали, что эффективность элонгации зависит от кодонного состава открытых рамок считывания (ОРС) [Varenne et al., 1984; Sorensen et al., 1989] и от вторичной структуры мРНК [Такуар et al., 2005; Tuller et al., 2011]. Однако, несмотря на огромное количество исследований, эта область остается недостаточно изученной, и предсказание эффективности трансляции мРНК у многих организмов является актуальной проблемой.

Важность оценки эффективности трансляции связана с таким понятием, как гетерологичная экспрессия [Welch et al., 2009 a, b]. Гетерологичной называется экспрессия чужеродного гена или искусственной генетической конструкции в целевом организме. В настоящее время известны структуры геномов большого числа одноклеточных организмов. Многие из них рассматриваются в качестве кандидатов для использования в биотехнологических процессах и экспериментах. Это часто требует экспрессии различных генетических конструкций в целевых организмах. Для максимизации эффективности гетерологичной экспрессии требуются знания о механизмах и факторах, ее определяющих, в том числе и знания об эффективности трансляции. Поэтому одной из актуальных задач современной биоинформатики является изучение различных характеристик мРНК, влияющих на эффективность трансляции. Кроме этого, сравнительный анализ трансляционно значимых параметров мРНК генов различных организмов ценен сам по себе, так как является источником информации об эволюционных аспектах формирования этих признаков, имеющих как универсальные для всех, так и видоспецифичные особенности.

Исследование контекстных характеристик ОРС, связанных с эффективностью трансляции, актуально как для одноклеточных, так и для многоклеточных организмов. Однако наличие у многоклеточных организмов тканеспецифичной экспрессии [Dittmar et al., 2006] не позволяет с достаточной точностью выявлять у них эти особенности.

В Институте цитологии и генетики был разработан математический индекс эффективности элонгации трансляции EEI (elongation efficiency index), позволяющий оценивать эффективность элонгации трансляции генов организма на основании их нуклеотидного состава [Лихошвай и Матушкин, 2000]. Данный индекс имеет смысл средней скорости движения рибосомы по мРНК в процессе элонгации трансляции. EEI учитывает кодонный состав ОРС и локальные совершенные инвертированные повторы (потенциальные вторичные структуры в мРНК). В зависимости от того, какие из этих факторов являются определяющими при оценке эффективности элонгации трансляции, у исследуемого организма определяется тип эволюционной оптимизации его генома для увеличения эффективности процесса элонгации трансляции генов.

### **Цели и задачи исследования**

Целью данной работы является: исследование контекстных характеристик открытых рамок считывания, связанных с эффективностью элонгации трансляции, у одноклеточных организмов. Для ее достижения были поставлены следующие задачи:

- 1) Разработать доступную через Интернет программную реализацию самообучающегося алгоритма расчета индекса эффективности элонгации трансляции EEI;
- 2) Классифицировать секвенированные геномы одноклеточных организмов по типам эволюционной оптимизации процесса элонгации трансляции;
- 3) Исследовать связанные с процессом трансляции особенности структурно-функциональной организации открытых рамок считывания у различных одноклеточных организмов;
- 4) Изучить взаимосвязь между эффективностью инициации транскрипции и эффективностью элонгации трансляции у *S. cerevisiae* и *S. pombe*.

### **Научная новизна**

Разработанное веб-приложение EloE (<http://www-bionet.sccc.ru:7780/EloE>) позволило впервые провести анализ полных геномов 2771 одноклеточного организма.

В результате анализа организмов, принадлежащих к роду *Mycoplasma*, у группы видов обнаружено сниженное количество локальных инвертированных повторов в генах по

сравнению с другими микоплазмами. Филогенетическое исследование *Mycoplasma* позволило установить возможную связь эволюционной оптимизации первичной структуры генов данных организмов с их средой обитания. Также было установлено наличие достоверной отрицательной корреляции между GC-составом генома и степенью эволюционной оптимизации первичной структуры генов для повышения эффективности элонгации трансляции. Показано, что *M. haemofelis*, возможно, обладает отличным от других микоплазм механизмом регуляции процесса инициации трансляции.

Анализ нуклеотидных последовательностей генов и их предковых форм у архей позволил установить, что наиболее сильные изменения в первичной структуре генов, связанные с оптимизацией элонгации трансляции, происходили при радикальной смене среды обитания данных организмов. Также для архей было показано, что температура среды обитания данных организмов не коррелирует с влиянием потенциальных вторичных структур в мРНК на эффективность элонгации трансляции.

При анализе генов дрожжей выявлено наличие корреляции между потенциалом формирования нуклеосом и индексом эффективности элонгации трансляции, что подтверждает предположение о согласованной оптимизации процессов транскрипции и трансляции. Обнаружено различие между *S. cerevisiae* и *S. pombe* по форме корреляции между потенциалом формирования нуклеосом и индексом эффективности элонгации трансляции для высоко- и низкоэкспрессирующихся генов. Проведен сравнительный анализ геномов этих организмов для выявления причин данного различия.

### **Теоретическая и практическая значимость**

Результаты данной работы могут быть использованы в генно-инженерных экспериментах для создания искусственных генетических конструкций. Оптимизация первичной структуры нуклеотидных последовательностей позволит увеличить эффективность их трансляции и тем самым повысить уровень их экспрессии в целевых организмах.

Также результаты могут быть полезны при работе с малоизученными организмами, для которых не доступны экспериментальные данные по экспрессии генов. Предсказанные уровни эффективности элонгации трансляции в первом приближении позволяют оценить эффективность экспрессии исследуемых генов.

В теоретическом плане данная работа содержит новую информацию по связанным с эффективностью элонгации трансляции особенностям геномов разнообразных организмов (архей, микоплазм, дрожжей). Эти знания могут послужить основой для проведения новых экспериментов или объяснения особенностей процесса трансляции.

### Положения, выносимые на защиту

- 1) У семи видов *Mycoplasma* (*C. M. haemolamae*, *M. haemocanis*, *M. wenyonii*, *M. haemofelis*, *M. pneumoniae*, *C. M. haemominutum*, *M. suis*), в процессе эволюции прошла массовая минимизация количества локальных совершенных инвертированных повторов (потенциальных шпилек) в мРНК.
- 2) *M. haemofelis* радикально отличается от остальных проанализированных видов микоплазм наличием более стабильных потенциальных вторичных структур в мРНК в районе старт-кодона трансляции, что может быть связано с альтернативным механизмом регуляции инициации трансляции у данного вида.
- 3) Индекс эффективности элонгации трансляции генов *S. cerevisiae* значимо коррелирует с экспериментально определенной плотностью нуклеосомной упаковки во фланкирующем 5'-районе ДНК выше старта трансляции мРНК.

### Апробация результатов

Данная работа была представлена на следующих конференциях:

- а) XIII всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям, Новосибирск, 2012;
- б) Moscow conference on computational molecular biology, MCCMB'13, Москва, 2013;
- в) 5th international young scientists school «Systems biology and bioinformatics», SBB'2013, Новосибирск, 2013;
- г) VI съезд Вавиловского общества генетиков и селекционеров (ВОГиС) и ассоциированные генетические симпозиумы, Ростов-на-Дону, 2014 (диплом 3-ей степени);
- д) The 9th international conference on bioinformatics of genome regulation and structure\System biology, BGRS\SB'2014, Новосибирск, 2014;
- е) 6th international young scientists school «Systems biology and bioinformatics», SBB'2014, Новосибирск, 2014.

### Личный вклад автора

Основные результаты работы получены автором самостоятельно. Разработка веб-приложения EtoE проводилась совместно с Б. С. Зураевым (создание веб-интерфейса, подключение программы UNAFold), к.б.н. С. А. Лашиным (консультации по коду программы), д.б.н. В. А. Лихошваем (консультации по алгоритмам программы) и к.б.н. Ю. Г. Матушкиным (консультации по алгоритмам программы). Исследование оптимизации первичной структуры

генов архей в процессе эволюции проводилось совместно с к.б.н. К. В. Гунбиным (реконструкция предковых форм генов архей). Исследование взаимосвязи между эффективностью элонгации трансляции генов дрожжей и плотностью их нуклеосомной упаковки в 5'-НТР проводилось совместно с к.б.н. В. Г. Левицким (расчет потенциала формирования нуклеосом), д.б.н. Ю. Л. Орловым (экспериментальные данные по плотности нуклеосомной упаковки), д.б.н. В. А. Лихошваем и к.б.н. Ю. Г. Матушкиным.

### Список работ, опубликованных по теме диссертации

#### Статьи:

- 1) Матушкин Ю. Г., Левицкий В. Г., **Соколов В. С.**, Лихошвай В. А., Орлов Ю. Л. Эффективность элонгации генов дрожжей коррелирует с плотностью нуклеосомной упаковки в 5'-нетранслируемом районе. Математическая биология и биоинформатика. 2013. 8(1):248–257.
- 2) **Sokolov V. S.**, Likhoshvai V. A., Matushkin Yu. G. Gene expression and secondary mRNA structures in different Mycoplasma species. Russian Journal of Genetics: Applied Research. 2014. 4(3):208–217.
- 3) **Sokolov V. S.**, Zuraev B. S., Lashin S. A., Matushkin Yu. G. EloE: web application for estimation of gene translation elongation efficiency. Russian Journal of Genetics: Applied Research. 2015. 5(4):335–339.
- 4) **Sokolov V. S.**, Zuraev B. S., Lashin S. A., Matushkin Yu. G. Web application for automatic prediction of gene translation elongation efficiency. Journal of Integrative Bioinformatics. 2015. 12(1):257–264.

#### Тезисы конференций:

- 1) **Соколов В. С.**, Лихошвай В. А., Матушкин Ю. Г. Программное обеспечение для компьютерного исследования особенностей элонгации трансляции (на примере одноклеточных организмов рода Mycoplasma). XIII всероссийская конференция молодых ученых по математическому моделированию и информационным технологиям, 2012. <http://conf.ict.nsc.ru/ym2012/ru/reportview/138699>.
- 2) **Соколов В. С.**, Матушкин Ю. Г. Компьютерное исследование особенностей элонгации трансляции у Mycoplasma. VI съезд Вавиловского общества генетиков и селекционеров (ВОГиС) и ассоциированные генетические симпозиумы, 2014. С. 64–65. Диплом 3-ей степени.

- 3) **Sokolov V. S.**, Likhoshvai V. A., Matushkin Yu. G. Gene expression and mRNA secondary structures in Mycoplasma strains. 5th international young scientists school «Systems biology and bioinformatics», SBB'2013. <http://conf.ict.nsc.ru/SBB2013/reportview/158983>.
- 4) **Sokolov V. S.**, Likhoshvai V. A., Matushkin Yu. G. Computational study of translation elongation features in Mycoplasma. Moscow conference on computational molecular biology, MCCMB'13. <http://mccmb.belozersky.msu.ru/2013/abstracts/abstracts/156.pdf>.
- 5) **Sokolov V. S.**, Gunbin K. V., Matushkin Yu. G. Variation of elongation efficiency index of Archaea genes during evolution. The 9th international conference on bioinformatics of genome regulation and structure\System biology, BGRS\SB'2014. P. 153.
- 6) **Sokolov V. S.**, Zuraev B. S., Lashin S. A., Matushkin Yu. G. EloE – web application for estimation of translation elongation efficiency of genes in various organisms. The 9th international conference on bioinformatics of genome regulation and structure\System biology, BGRS\SB'2014. P. 152.
- 7) **Sokolov V. S.**, Matushkin Yu. G. Analysis of Bacteria and Archaea genomes available in GenBank database by “EloE” program. 6th international young scientists school «Systems biology and bioinformatics», SBB'2014. P. 33.

#### **Авторское свидетельство**

**Соколов В. С.**, Зураев Б. С., Генаев М. А. «Программа для автоматической оценки эффективности элонгации трансляции генов различных организмов (EloE)», № 2014662021 от 19.11.2014.

#### **Структура и объем работы**

Диссертационная работа состоит из списка сокращений, введения, обзора литературы, методов и алгоритмов, результатов и обсуждений, заключения, выводов, списка литературы и приложения. Работа изложена на 163 страницах, содержит 74 рисунка и 9 таблиц. Библиографический указатель литературы включает 196 источников, из них 2 отечественных и 194 зарубежных.

#### **Благодарности**

Автор выражает глубокую признательность сотрудникам лаборатории молекулярно-генетических систем ИЦиГ СО РАН, и лично заведующему лабораторией и научному руководителю к.б.н. Матушкину Ю. Г., а также соавторам и коллегам – академику РАН Колчанову Н.А., д.б.н. Лихошваю В.А., к.б.н. Гунбину К.В., к.б.н. Левицкому В.Г. за



консультации и плодотворные научные дискуссии. Автор благодарит д.б.н. Кочетова А.В. и к.б.н. Афонникова Д.А. за полезные критические замечания на стадии подготовки диссертации. Кроме того, автор выражает благодарность к.б.н. Лашину С.А. (Лаборатория молекулярно-генетических систем, ИЦиГ СО РАН) за квалифицированную помощь и консультации при написании программ на языке Java.

## ГЛАВА 1. ОБЗОР ЛИТЕРАТУРЫ

### 1.1. Механизм процесса трансляции на примере трансляции *Escherichia coli*

Процесс трансляции различается (хотя и не очень сильно) в прокариотах, археях и эукариотах. В основном отличия касаются процессов инициации и терминации. Результаты, описываемые в данной работе, получены на одноклеточных организмах (в том числе эукариотических). Поэтому в обзоре литературы мы рассматриваем в качестве модельного процесс трансляции в *E. coli*, как хорошо изученном модельном организме.

Процесс трансляции состоит из трех основных стадий: инициации, элонгации и терминации.

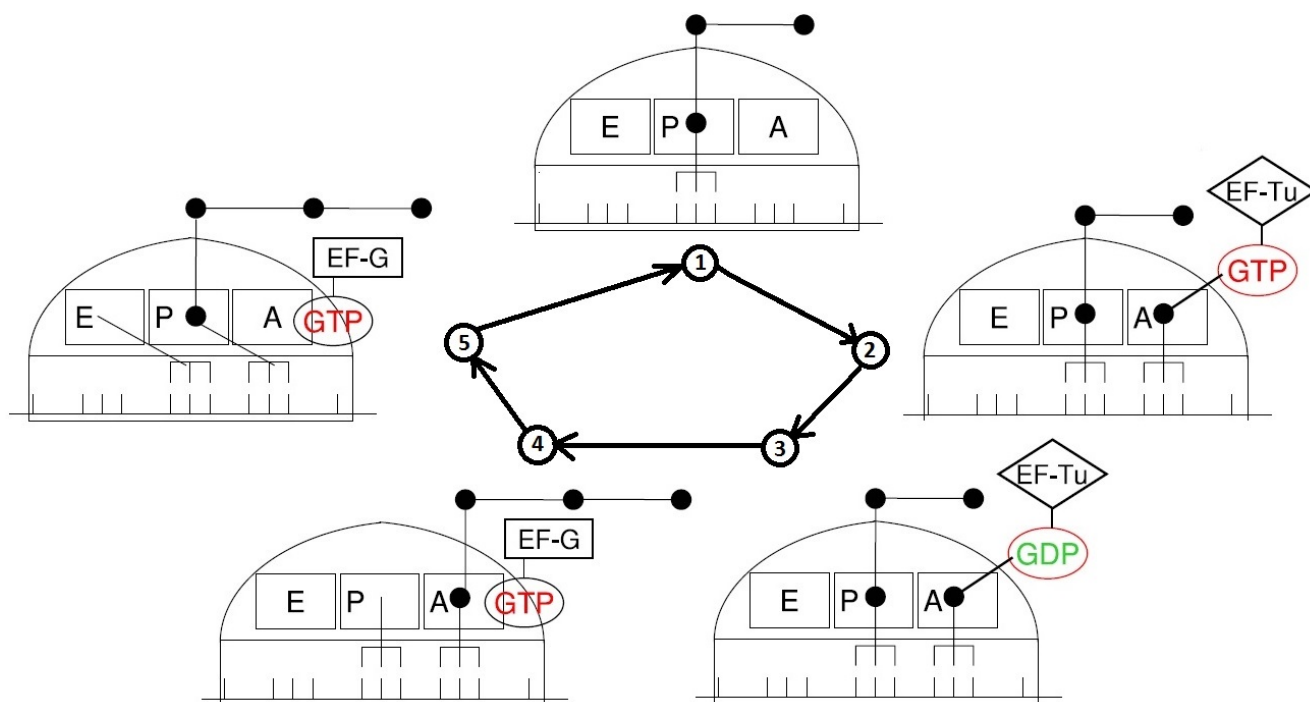
**Инициация.** Данная стадия обеспечивается тремя факторами инициации: IF1, IF2 и IF3. Они регулируют взаимодействие между последовательностью Шайна-Дальгарно (ШД) на мРНК (примерно 6-9 нуклеотидов выше старт-кодона трансляции) и последовательностью анти-ШД на 3'-конце 16S рРНК [Kaczanowska and Rydén-Aulin, 2007].

В соответствии с Шайном и Дальгарно, 5' не транслируемый район (5'-НТР) *E. coli* содержит нуклеотидную последовательность, соответствующую консенсусу AGGAGG. Данная последовательность комплементарна последовательности на 3'-конце 16S рРНК. В работе [Osada et al., 1999] было рассчитано значение свободной энергии взаимодействия между данными последовательностями у различных видов прокариот, эукариот и архей. В результате было показано, что среднее значение этой энергии значительно снижено в районе перед стар-кодоном трансляции у *E. coli*, что подтверждает наличие связи между 5'-НТР гена и 16S рРНК в данном организме. У *Haemophilus influenza*, *B. subtilis*, *H. pylori* и *A. aeolicus* были показаны аналогичные результаты. Это говорит о том, что все эти организмы имеют схожий механизм инициации трансляции.

Другие бактерии (*Synechocystis* PCC6803, *M. genitalium*, *M. pneumoniae* и *B. burgdorferi*) также характеризуются снижением значения свободной энергии взаимодействия между последовательностями ШД и анти-ШД. Но этот спад менее значителен и сильно отличается от такового у *E. coli*. Аналогичный анализ был проведен для геномов эукариот – снижения значения свободной энергии взаимодействия между нуклеотидными последовательностями на 3'-конце 18S рРНК и 5'-НТР у *S. cerevisiae* обнаружено не было. Графики зависимости значений свободной энергии от позиции для трех видов архей (*A. fulgidus*, *M. jannaschii* и *M. thermoautotrophicum*) похожи на бактериальные и отличаются от *S. cerevisiae*. Это говорит о том, что археи ближе к бактериям, чем к эукариотам в отношении механизма инициации трансляции [Osada et al., 1999].

После ассоциации мРНК с 30S субъединицей рибосомы, факторы инициации способствуют расположению старт-кодона трансляции в Р-сайте рибосомы. Далее происходит присоединение формилметионин-тРНК (fMET-тРНК) и 50S субъединицы рибосомы. Стадия инициация считается лимитирующим звеном трансляции [Kaczanowska and Rydén-Aulin, 2007].

**Элонгация.** В начале данной стадии Р-сайт рибосомы содержит fMET-тРНК, а А-сайт пуст и готов к принятию очередной аминокислотированной тРНК (Рисунок 1, 1). Фактор элонгации EF-Tu способствует расположению очередной тРНК в А-сайте рибосомы, в процессе чего происходит гидролиз ГТФ (Рисунок 1, 2-3). В результате, концы тРНК, расположенных в Р и А сайтах рибосомы, оказываются в пептидил-трансферазном центре (ПТЦ) 50S субъединицы рибосомы. Образуется пептидная связь между аминокислотами, и полученный полипептид оказывается связанным с тРНК, находящейся в А-сайте (Рисунок 1, 4). На следующем этапе деацилированная тРНК перемещается из Р-сайта в Е-сайт для дальнейшего выхода из рибосомы, а пептидил-тРНК – из А- в Р-сайт (Рисунок 1, 5). Данный процесс называется транслокацией и обеспечивается фактором EF-G и гидролизом ГТФ. В результате рибосома смещается на один кодон вдоль мРНК, после чего готова к повторению данного цикла (Рисунок 1, 1) [Kaczanowska and Rydén-Aulin, 2007].



**Рис. 1.** Основные этапы трансляции (см. текст). GTP – гуанозинтрифосфат; GDP – гуанозиндифосфат; EF-Tu – фактор элонгации трансляции; EF-G – факторы транслокации трансляции; А, Р и Е – активные центры рибосомы. Фрагменты рисунка взяты из статьи [Sharma and Chowdhury, 2011].

**Терминация.** Данная стадия начинается с попадания в А-сайт рибосомы стоп-кодона трансляции. Данный кодон распознается факторами терминации RF1 и RF2. RF1 распознает UAA и UAG кодоны, а RF2 – UAA и UGA. Присоединение RF1/RF2 к рибосоме инициирует

гидролиз и высвобождение пептидной цепи от тРНК в Р-сайте. В это время третий фактор RF3 связывается с рибосомой и способствует диссоциации RF1/RF2 из А-сайта. После этого под действием различных факторов происходит отсоединение рибосомы от мРНК и деацилированной тРНК от рибосомы [Kaczanowska and Rydén-Aulin, 2007].

У *E.coli* и других прокариот известно такое явление как сопряженная трансляция (реинициация трансляции) [Platt et al., 1972; Ganem et al., 1973]. Это процесс повторного вступления рибосом, терминировавших биосинтез белка с одного гена, в цикл трансляции следующего за ним гена без предварительного отсоединения от мРНК. Опероны у *E.coli* – это расположенные друг за другом на ДНК гены (цистроны), которые транслируются совместно. Реинициация широко распространена у *E. coli* и играет важную роль в контроле экспрессии генов на уровне трансляции.

## **1.2. Контекстные особенности открытых рамок считывания, связанные с эффективностью элонгации трансляции**

На эффективность элонгации трансляции гена оказывают влияние множество факторов. Наиболее подробно в данном разделе рассмотрены такие факторы как кодонный состав открытых рамок считывания и наличие вторичных структур в мРНК.

### **1.2.1. Влияние кодонного состава открытых рамок считывания на эффективность элонгации трансляции**

#### **1.2.1.1. Неравномерность использования синонимичных кодонов в открытых рамках считывания**

Известно, что генетический код является вырожденным, т.е. одну аминокислоту в белке могут кодировать несколько синонимичных кодонов [Hershberg and Petrov, 2008]. Поскольку синонимичные замены в генах не приводят к изменениям в первичной структуре белка, считалось, что синонимичные кодоны должны использоваться при кодировании примерно с одинаковой частотой, равновероятно. Однако по мере появления все большего количества данных секвенирования стало очевидно, что синонимичные кодоны в генах различных организмов имеют разную частоту встречаемости. Наиболее часто встречаемые синонимичные кодоны называют предпочтительными.

Grantham и его коллеги обнаружили, что синонимичные кодоны используются с разной частотой в разных организмах [Grantham 1980; Grantham et al., 1980, 1981]. Этот факт был назван «геномной гипотезой». В дальнейшем было установлено, что по частотам кодонов различаются не только организмы, но и сами гены внутри одного организма. Например, было

показано, что предпочтительные кодоны для генов митохондрий дрожжей значительно отличаются от аналогичных для ядерных генов [Bonitz et al., 1980]. Была установлена связь между степенью неравномерности использования кодонов (НИК) в гене с его уровнем экспрессии. Под неравномерностью подразумевается разная частота встречаемости синонимичных кодонов в гене.

Неравномерность в использовании синонимичных кодонов при кодировании аминокислот в белках была обнаружена во многих организмах [Grantham et al., 1980; Sharp and Li, 1987; Andersson and Kurland, 1990; Wada et al., 1990; Stenico et al., 1994]. Стало ясно, что набор кодонов, специфичный для конкретного организма, связан со специфичным набором изоакцепторных тРНК в этом организме, по крайней мере, в случае *E. coli* и дрожжей [Welch et al., 2009b]. В нескольких работах была показана корреляция между частотами кодонов и концентрациями соответствующих им молекул тРНК [Bennetzen and Hall, 1982; Gouy and Gautier, 1982; Ikemura, 1985]. Чем больше в мРНК предпочтительных кодонов, тем быстрее проходит стадия элонгации трансляции для данного гена, т.к. не происходит задержки рибосомы на кодонах, которым соответствуют тРНК с низкой концентрацией [Varenne et al., 1984; Sorensen et al., 1989]. По этой причине предпочтительные кодоны также называют оптимальными.

Одними из первых организмов, для которых было установлено неравномерное (смещенное в сторону предпочтительных) использование кодонов, были *E. coli* и *S. cerevisiae*. У *S. cerevisiae* обнаружены гены с экстремальным смещением кодонов: алкогольдегидрогеназа I (ADH-I) и глицеральдегид-3-фосфат дегидрогеназа. 96% из 1004 аминокислот кодируются 25 триплетами из 61 возможного. Эти кодоны гомологичны антикодонам наиболее распространенных видов изоакцепторных тРНК. Везде, где это возможно, в генах *S. cerevisiae* избегается использование кодонов, в которых друг за другом располагаются GC пары оснований. Также избегаются мононуклеотидные кодоны (GGG, CCC, AAA, UUU) и кодоны, состоящие только из GC или AU. Это приводит к приблизительно равной энергии кодон-антикодонной связи для всех предпочтительных кодонов. Количество предпочтительных кодонов в генах *S. cerevisiae* варьирует от 20% (у изо-2-цитохрома C) до более чем 90%. Степень смещения кодонов для каждого гена положительно коррелирует с количеством соответствующей мРНК в цитоплазме. Аналогичная корреляция наблюдается у *E. coli*, только предпочтительные кодоны в данном организме другие. [Bennetzen and Hall, 1982].

Частотный спектр используемых кодонов в организме формирует его «диалект». Исследование «диалектов» различных одноклеточных организмов показало, что «диалект» *E. coli* похож на «диалекты» других организмов из семейства Enterobacteriaceae (например, *Shigella*, *Salmonella*, *Klebsiella*, *Serratia*, и *Erwinia*) и отличается от таксономически более

далеких (например, *Anabaena* и *Bacillus*). Это говорит о том, что популяции молекул тРНК достаточно консервативны в процессе эволюции [Nichols et al., 1981].

Yanofsky и его коллеги показали, что выбор кодонов в организмах с повышенным GC-составом частично определяется именно им [Yanofsky and van Cleemput, 1982].

### 1.2.1.2. Молекулярные механизмы и математические модели, объясняющие неравномерность использования кодонов и ее корреляцию с пулом тРНК

Ikemura [Ikemura and Ozeki 1983] предложил следующий молекулярный механизм того, как количество определенных изоакцепторных тРНК влияет на выбор кодонов в процессе эволюции. В процессе синтеза белка затрачивается большое количество энергии и вещества. В случае *E. coli* 70% всей энергии клетки расходуется на этот процесс, а масса всех рибосом составляет примерно 1/3 сухой массы клетки. Икемура предположил, что кодонный «диалект» должен быть отражением стратегии организма по синтезу большого количества белка с минимальными затратами [Ikemura and Ozeki 1983]. Если высокотранслируемые гены будут содержать кодоны, соответствующие редким тРНК, это будет приводить к более частому попаданию в А-сайт рибосомы некомплементарных тРНК (тРНК с некомплементарным антикодоном). Это в свою очередь приведет к затратам энергии на то, чтобы заменить некомплементарную тРНК комплементарной. Таким образом, случайные мутации в синонимичных позициях кодонов высокоэкспрессируемых генов, приводящие в соответствие кодону распространенную изоакцепторную тРНК, будут снижать энергетические затраты клетки на процесс трансляции данного гена. Также это приведет к ускорению трансляции остальных генов, так как освободит рибосомы, которые раньше задерживались на медленных кодонах.

В 1987 году Michael Bulmer предложил математическую модель, которая, несмотря на ее простоту, позволила описать коэволюцию частот использования кодонов и количества изоакцепторных тРНК [Bulmer, 1987].

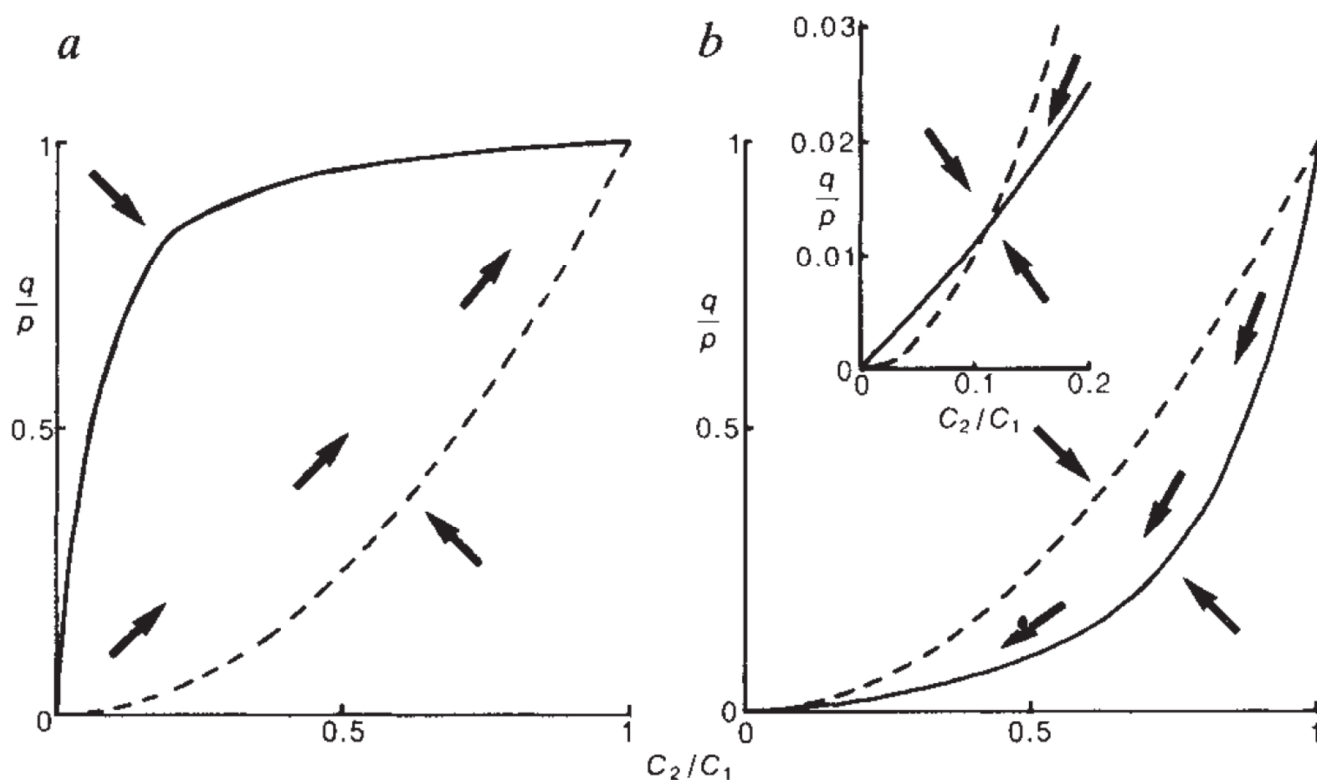
$$q = \frac{1}{2} + \frac{u}{s} - \frac{1}{2} \left[ 4 \left( \frac{u}{s} \right)^2 + 1 \right]^{1/2} \quad (1)$$

$$s = \alpha(t_2 - t_1) = k \left( \frac{C_1}{C_2} - 1 \right) \quad (2)$$

$$k = \alpha\theta / C_1 \quad (3)$$

Где  $q$  – частота кодона,  $u$  – скорость мутации из одного кодона в другой,  $s$  – приспособленность кодона,  $\theta$  – среднее время распознавания комплементарной тРНК рибосомой,  $C_i$  – концентрация  $i$ -ой тРНК,  $\alpha$  – параметр, зависящий от эффективности экспрессии гена.

Первая часть модели (Формула 1) соответствует предположению, что в процессе эволюции изменяются только частоты кодонов, а количество изоакцепторных тРНК остается постоянным. Это позволяет показать, как прямое влияние времени трансляции на приспособленность организма может обеспечить достаточное давление отбора, чтобы объяснить НИК. Вторая часть (Формулы 2-3), наоборот, соответствует предположению о постоянстве частот кодонов и эволюцию пула тРНК.



**Рис. 2.** Предсказанное использование кодонов, изображенное относительно количества тРНК для (а) низкоэкспрессирующихся и (б) высокоэкспрессирующихся генов. Вставка в (b) – увеличенный участок графика вблизи нуля. Сплошная линия – эволюция использования кодонов при фиксированном количестве тРНК. Пунктирная линия – эволюция количества тРНК ( $C_1$  и  $C_2$ ) при фиксированных частотах кодонов ( $p$  и  $q$ ). Стрелки показывают направление движения. (а) точка неустойчивого равновесия – (0, 0); точка устойчивого равновесия – (1; 1). (б) точки неустойчивого равновесия: (0; 0) и (1; 1); точка устойчивого равновесия – (0,113; 0,013). Ось абсцисс – отношение количества тРНК; ось ординат – отношение частот кодонов. Рисунок из статьи [Bulmer, 1987].

Объединение результатов обеих частей этой модели позволило исследовать коэволюцию частот кодонов и количества изоакцепторных тРНК. Результаты показали, что в низкоэкспрессирующихся генах у модели есть одна точка устойчивого равновесия (1; 1) (Рисунок 2, а), к которой стремится система. Она характеризуется небольшой неравномерностью по использованию кодонов в гене и примерно равными количествами изоакцепторных тРНК. В высокоэкспрессируемых генах модель также имеет одну точку равновесия (0,113; 0,013) (Рисунок 2, б; увеличенная область), но она отличается от

предыдущей. Данное положение равновесия характеризуется неравным составом пула тРНК и высокой неравномерностью кодонного состава генов [Bulmer, 1987].

Результаты данной работы позволяют объяснить наблюдаемую высокую неравномерность использования кодонов, ассоциированную с неоднородным пулом тРНК, у одноклеточных организмов. Для них высокая скорость деления является важным фактором приспособленности, из-за чего появляется высокое давление отбора для минимизации времени трансляции. Чтобы получить НИК в рамках данной модели, среднее давление отбора на нуклеотид в ответ на неравномерность пула тРНК должно превышать как скорость мутации одного нуклеотида, так и величину, обратную эффективной численности популяции. Иначе неравномерность кодонов будет определяться лишь мутационными процессами. Данные показывают, что эти условия выполняются для *E. coli* и *S. cerevisiae* вкупе с их большими размерами популяций. С другой стороны, многоклеточные организмы характеризуются гораздо меньшими размерами популяций и, по-видимому, низким давлением отбора на уровне трансляции. Это приводит к тому, что частоты кодонов в их генах в основном определяются мутационным отбором [Bulmer, 1987].

Другие исследования показали, что выбор среди синонимичных кодонов у некоторых организмов определяется не только из соображений оптимизации трансляции. Движущими силами отбора могут быть такие факторы как: специфичная для каждой из цепей ДНК неравномерность мутаций; естественный отбор на уровне не только трансляции, но и репликации и транскрипции; уровень гидрофобности белка [Romero et al., 2000; Ermolaeva, 2001].

Показано, что среди прокариот влияние естественного отбора и мутационной неравномерности различны в зависимости от того, является ли геном организма АТ- или GC-богатым. В организмах с низким GC-составом преобладает влияние мутационной неравномерности на кодонный состав. С другой стороны, в организмах с высокими значениями GC-состава наибольшую роль играет именно трансляционный отбор [Naya et al., 2001].

В 2001 году впервые было показано, что частоты кодонов определяются трансляционным отбором в одноклеточных эукариотах с экстремально высоким GC-составом. Был сделан вывод, что в условиях большого эффективного размера популяции и короткого времени размножения для одноклеточных эукариот трансляционный отбор всегда работает, перекрывая эффекты сильного мутационного смещения [Naya et al., 2001].

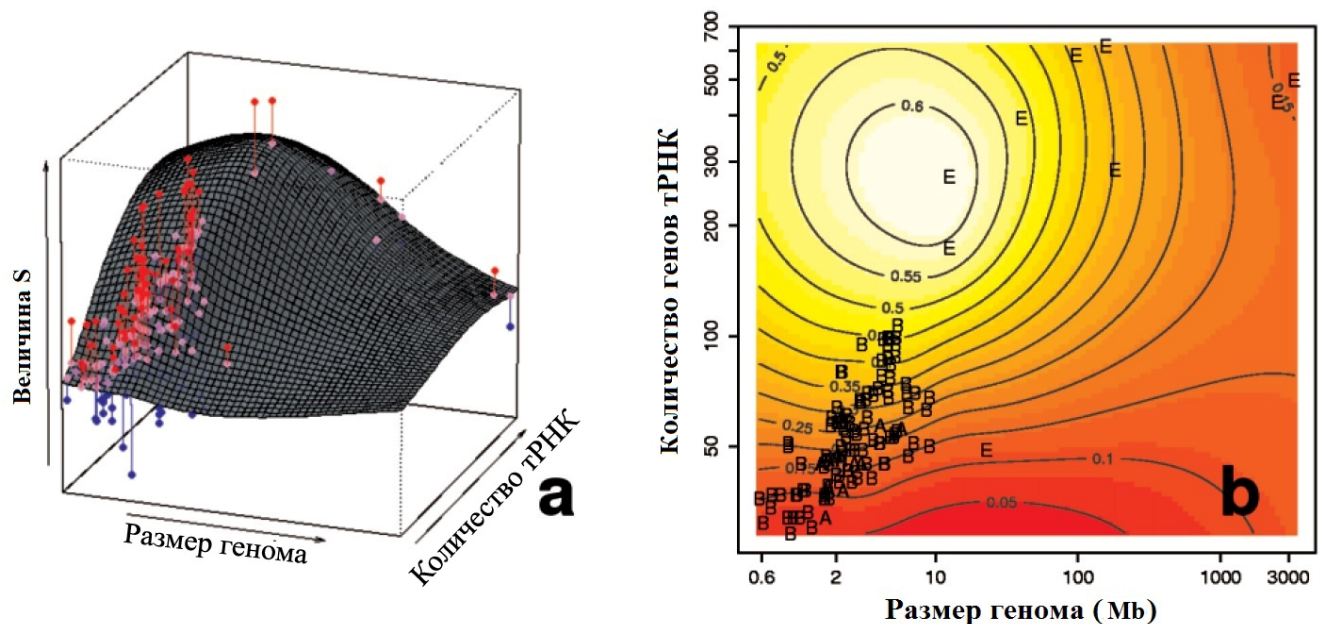
Как уже было сказано, трансляционный отбор ответственен за неравномерное использование синонимичных кодонов в генах, кодирующих белки, в большом разнообразии организмов. В статье [Reis et al., 2004] была предложена мера S, позволяющая оценить вклад трансляционного отбора в НИК, без учета вклада GC-состава.



Мера  $S$  представляет из себя коэффициент корреляции между векторами  $tA_{IG}$  и  $\Psi_G$ .  $tA_{IG}$  – мера адаптированности гена к пулу тРНК организма (подробнее см. раздел 1.3.1. главы «Обзор литературы»).  $\Psi_G = \phi_G - \varepsilon_G$ , где  $\phi_G$  – давление отбора на используемые кодоны,  $\varepsilon_G$  – влияние случайных факторов на использование кодонов [Reis et al., 2004].

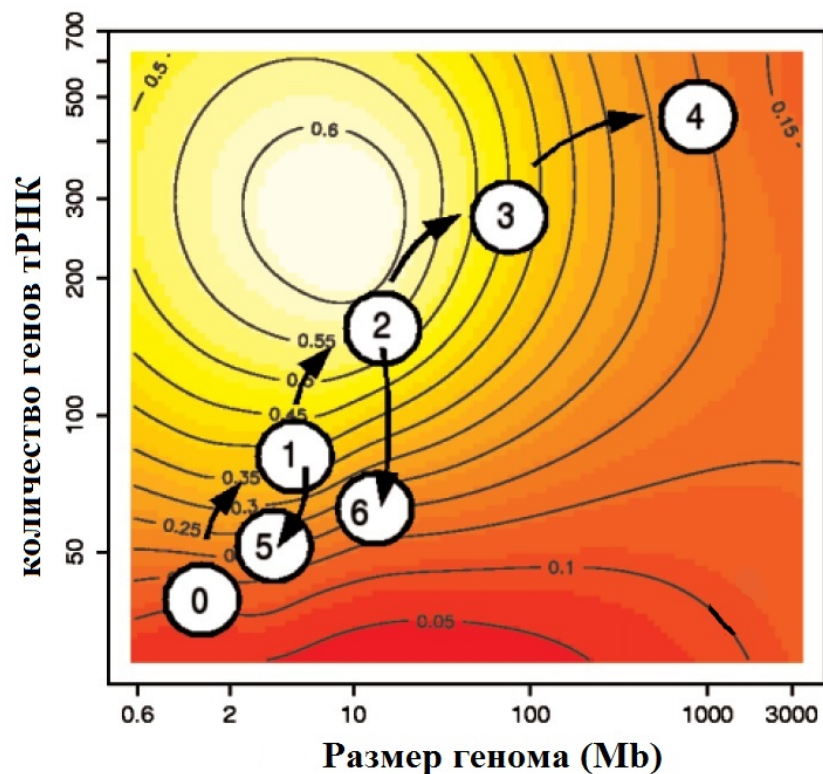
При помощи меры  $S$  был проведен анализ 126 полных геномов различных видов организмов от архей до эукариот. В результате было показано, что количество копий генов тРНК и размер генома влияют на действие трансляционного отбора, и что существует оптимальный размер генома, для которого этот тип отбора максимален. Соответственно, размер генома имеет верхнюю и нижнюю границы, за которыми отбор по использованию кодонов невозможен. Эти параметры формируют ландшафт, который отображает силу трансляционного отбора (Рисунок 3). Поверхность имеет максимум, где давление отбора на эволюцию синонимичных кодонов максимально, и области с низким уровнем отбора.

Маленькие бактериальные геномы (такие как у *H. pylori* или *B. burgdorferi*) и большие эукариотические геномы (*H. sapiens* или *M. musculus*) попадают в области с низким давлением отбора. Геномы дрожжей попадают в область максимума. В предложенной модели коэволюция размера генома и количества генов тРНК объясняет наблюдаемые закономерности в трансляционном отборе в различных организмах. Эта модель окончательно унифицирует наше понимание использования кодонов у прокариот и эукариот. Данная модель позволяет лучше понять особенности использования кодонов в *H. pylori*, *S. cerevisiae* и *H. sapiens*, как характерных представителей своих групп [Reis et al., 2004].



**Рис. 3.** Сила трансляционного отбора используемых кодонов в геномном ландшафте. (a) Поверхность регрессии зависимости силы отбора (мера  $S$ ) от количества копий генов тРНК и размера генома. (b) Цветовое изображение этой же поверхности. Рисунок по статье [Reis et al., 2004].

В рамках работы, проведенной Reis и коллегами, также сделано предположение о возможном пути эволюции по оптимизации кодонного состава (Рисунок 4). Гипотетический предковый организм обладал маленьким геномом и небольшим набором генов тРНК (Рисунок 4, 0). После нескольких последовательных стадий увеличения генома у него возросло количество генов тРНК (Рисунок 4, 1). Это привело к избыточному набору тРНК и, соответственно, к увеличению давления отбора на кодонный состав. На данном этапе появились бактерии со средним размером генома, такие как *E. coli*. Дальнейшее увеличение могло привести к появлению первых эукариотических геномов, таких как у дрожжей, у которых оптимизация кодонов наиболее развита (Рисунок 4, 2). С дальнейшим увеличением размеров геномов другие экологические составляющие начали препятствовать действию отбора на кодонный состав (Рисунок 4, 3). В этот момент появились большие современные геномы, такие как у млекопитающих (Рисунок 4, 4). Отбор на уменьшение размеров геномов или набора тРНК в определенных не свободно живущих организмах (например, *H. pylori* или *P. falciparum*) мог развернуть этот процесс в обратную сторону (Рисунок 4, 5-6) [Reis et al., 2004].



**Рис. 4.** Гипотетический путь эволюции оптимизации кодонного состава. Предок с маленьким геномом (0) прошел путь из нескольких увеличений генома (1-4). В течение этого эволюционного процесса давление отбора на кодонный состав сначала усиливалось, а затем ослабевало. Уменьшение размера генома или количества копий генов тРНК у паразитических организмов могло обратить данный процесс, например, *H. pylori* (5) или *P. falciparum* (6). Рисунок по статье [Reis et al., 2004].

Однако данная модель все равно не является универсальной и требует дальнейшего изучения. Например, некоторые бактерии с экстремальными значениями GC-состава

характеризуются околонулевыми значениями силы трансляционного отбора, хотя и обладают средним размером генома и средним количеством генов тРНК (*Clostridium tetani*: GC = 0,29; S = 0,03) [Reis et al., 2004].

### **1.2.1.3. Связь функции гена с адаптацией его кодонного состава к пулу тРНК**

В результате анализа ~2800 ортологичных генов в девяти геномах дрожжей было установлено, что на фенотипические различия между видами оказывает влияние не только эффективность транскрипции, но и трансляции. Показано, что ортологичные гены по-разному адаптируют свой кодонный состав к похожим пулам тРНК у разных видов дрожжей в зависимости от нужности этих генов в каждом конкретном организме. Например, гены, имеющие отношение к сплайсингу, имеют наибольшую приспособленность к пулу тРНК (наибольшую теоретическую эффективность трансляции) у *A. nidulans* и *S. pombe*. Именно у данных видов наблюдается наибольшее относительное количество генов, содержащих интроны (85% и 50% соответственно). Для сравнения у *S. cerevisiae* только 5% генов содержат интроны. Также у *Y. lipolytica* гены, связанные с метаболизмом органических кислот, имеют наибольшую эффективность трансляции по сравнению с другими видами. Это согласуется с фактом использования данного вида дрожжей для промышленного производства органических кислот. Таким образом, кодонный состав, по-видимому, находится под сильным давлением отбора. Это приводит к разной адаптации кодонов к пулам тРНК у различных видов в зависимости от функции гена и потребности в нем организма [Man and Pilpel, 2007].

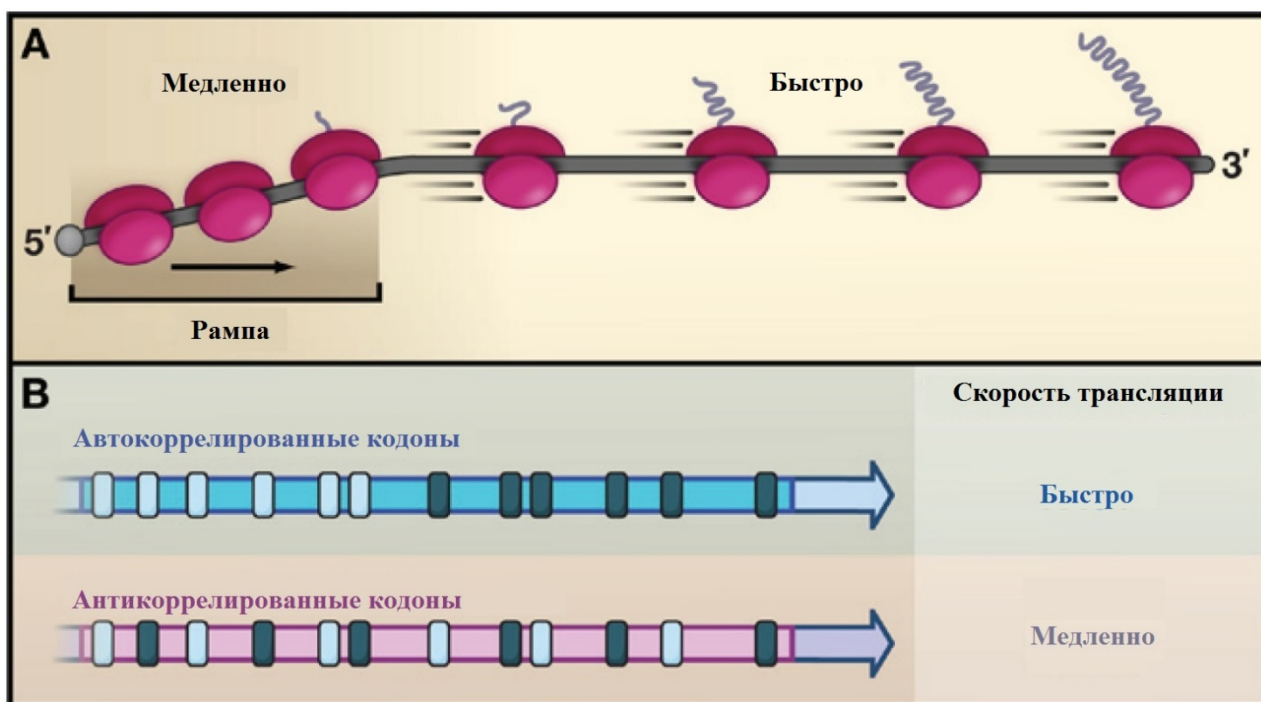
### **1.2.1.4. Медленные кодоны в 5'-районе кодирующей части генов. Модель взаимодействия рибосомы с тРНК**

Как было показано ранее, кодоны, соответствующие тРНК с высокой концентрацией в клетке, транслируются быстрее кодонов, соответствующих тРНК с низкой концентрацией. Такие кодоны называются быстрыми и медленными соответственно.

Новые экспериментальные данные по трансляции у дрожжей были получены с использованием метода рибосомного профилирования. Он основан на глубоком секвенировании мРНК фрагментов, связанных с рибосомами (футпринты), и позволяет проводить полногеномные исследования трансляции с субкодонным разрешением. Эти данные позволили показать, что эффективность трансляции гена значительно влияет на уровень его экспрессии, наряду с эффективностью транскрипции. Количество белка, определенное методом масс-спектрометрии протеома, лучше коррелирует с измеренной скоростью трансляции ( $r = 0,42$ ), чем с количеством мРНК ( $r = 0,17$ ) [Ingolia et al., 2009].

Описанная выше методика позволила оценить распределение рибосом по всей длине мРНК. Была рассчитана их плотность – среднее число рибосом на единицу длины. Показано, что средняя плотность рибосом значительно выше (примерно в три раза) на первых 30-40 кодонах, по сравнению с остальной частью гена, и снижается до среднего уровня после 100-200 кодонов ниже старт-кодона трансляции. Учет этого эффекта позволил повысить значение рассчитанного ранее коэффициента корреляции между измеренной скоростью трансляции и количеством белка до  $r = 0,60$ . Это говорит о том, что снижение плотности рибосом вдоль транскрипта связано с увеличением скорости трансляции [Ingolia et al., 2009].

Считается, что наличие в начале гена редких кодонов, замедляющих движение рибосом (так называемая «рампа»), способствует снижению вероятности сближения рибосом вплотную друг к другу при дальнейшем движении по мРНК. Такое сближение может приводить к полной остановке рибосом и даже ранней терминации трансляции. Таким образом, рампа в начале гена может способствовать повышению скорости трансляции (Рисунок 5, А) [Fredrick and Ibba, 2010].



**Рис. 5. (А)** Для многих генов установлено наличие кодонов, соответствующих редким тРНК, в их начале. Это подтверждает то, что скорость трансляции на протяжении примерно 50 первых кодонов ниже, чем на остальной части мРНК. **(В)** Расположение синонимичных кодонов вдоль гена влияет на скорость трансляции. Представлен простой пример, в котором два различных кодона кодирует одну и ту же аминокислоту. Когда одинаковые кодоны расположены последовательно вдоль мРНК (автокоррелировано), трансляция идет быстрее, чем когда они расположены вперемешку (антикоррелировано). Рисунок по статье [Fredrick and Ibba, 2010].

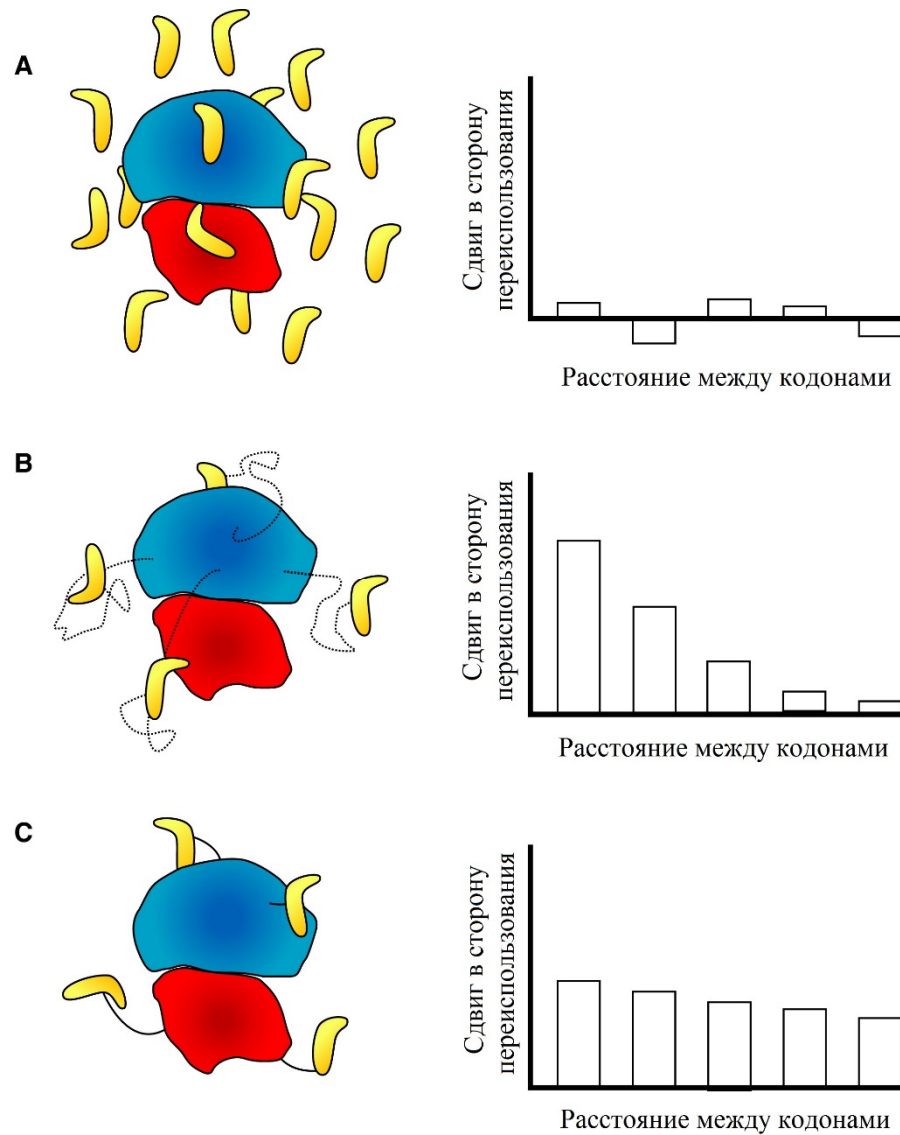
Однако это объяснение не единственное. Длина ramпы также хорошо согласуется с длиной полипептида, необходимого для заполнения выходного туннеля у рибосомы. Кроме

этого возможно, что медленная трансляция первых 20-30 кодонов способствует корректному взаимодействию между синтезируемым пептидом и шаперонами, тем самым увеличивая долю правильно свернутого белка. Однако, оба этих объяснения требуют экспериментального подтверждения [Tuller et al., 2010].

При исследовании распределения пар синонимичных кодонов в кодирующих последовательностях генома дрожжей была установлена зависимость скорости трансляции от взаимного расположения кодонов, кодирующих одну и ту же аминокислоту. Если две ближайшие одинаковые аминокислоты кодируются одинаковыми кодонами, такое расположение называется автокоррелированным, а если разными – антикоррелированным (Рисунок 5, В). Оказалось, что у генов *S. cerevisiae* скорость трансляции выше, если кодоны (быстрые или медленные) автокоррелированы. Такой же анализ был проведен для других эукариот (*Arabidopsis thaliana*, *Ashbya gossypii*, *Caenorhabditis elegans*, *Candida glabrata*, *Drosophila melanogaster*, *Homo sapiens*, и *S. pombe*). Все, кроме *S. pombe*, показали аналогичные результаты. Также установлено, что данный эффект ослабляется при увеличении расстояния между синонимичными кодонами [Fredrick and Ibba, 2010].

Таким образом, последовательности, которые поддерживают переиспользование тРНК, транслируются более эффективно, чем последовательности, которые поддерживают смену тРНК. К этому заключению приводят пять аргументов: 1) последовательности, различающиеся только порядком кодонов, а не частотами или кодируемыми аминокислотами, транслируются быстрее при условии автокоррелированности; 2) все исследованные геномы имеют смещение в сторону автокоррелированных последовательностей; 3) наиболее сильная автокорреляция наблюдается в высокоэкспрессируемых генах; 4) коррелированность сильнее для редких кодонов, особенно в высокоэкспрессируемых генах, что говорит о значительном влиянии автокорреляции на процесс трансляции; 5) данный эффект уменьшается с увеличением расстояния между синонимичными кодонами, демонстрируя эффект памяти во время трансляции [Cannarozzi et al., 2010].

Указанные данные позволяют предположить, что диффузия тРНК от рибосомы происходит медленнее чем трансляция и ацилирование. Это говорит о том, что ацилирование не является лимитирующим звеном. Было рассмотрено три модели взаимодействия тРНК с рибосомой (Рисунок 6). Модель А описывает быструю диффузию (быстрее трансляции) тРНК от рибосом. Модель В описывает более медленную диффузию. Модель С рассматривает физическое взаимодействие между тРНК и рибосомой. Из экспериментальных данных следует, что модель С наиболее вероятна [Cannarozzi et al., 2010].



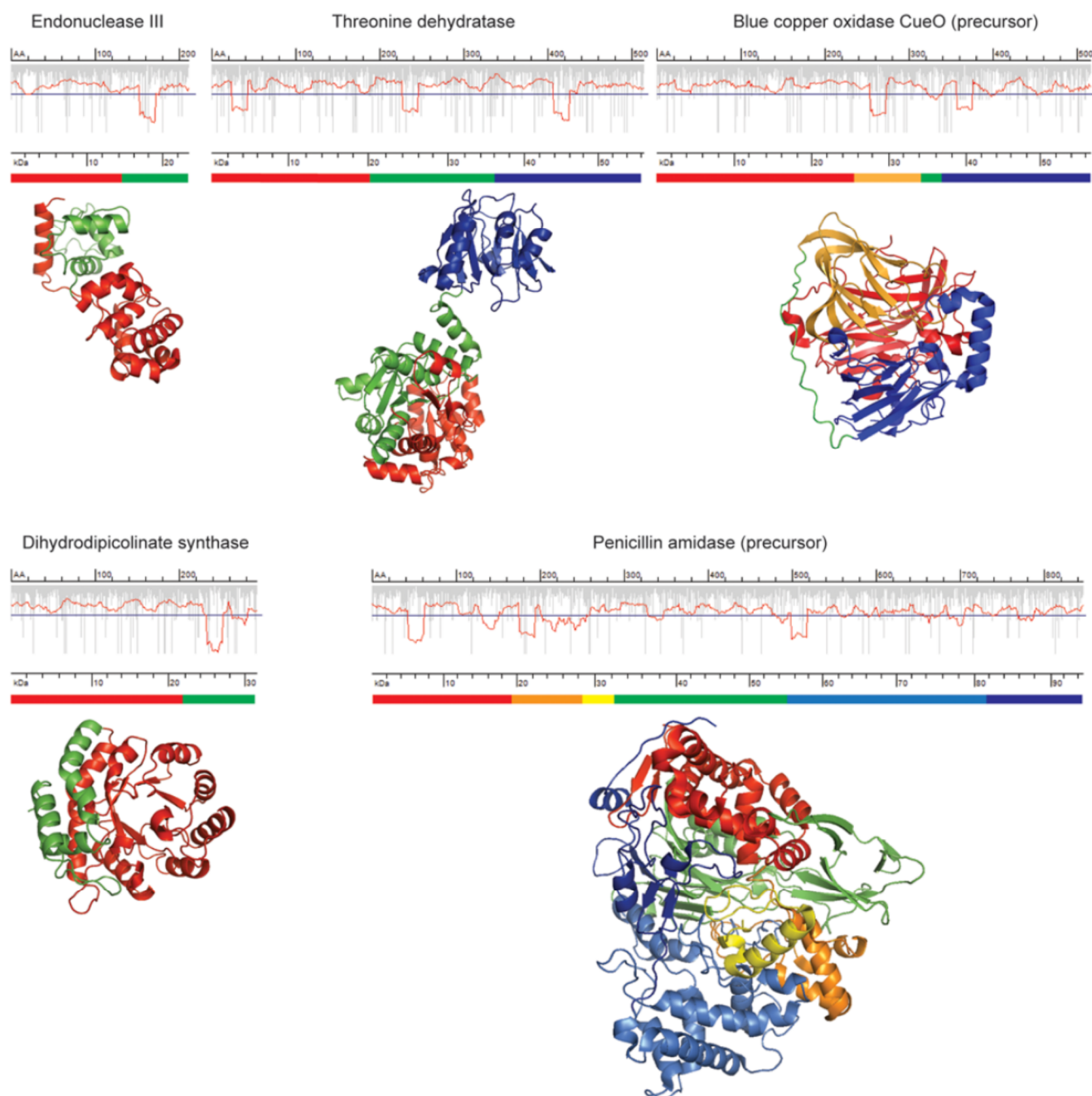
**Рис. 6.** Три гипотезы для объяснения коррелированности кодонов. Размеры и формы рибосом и тРНК адаптированы из кристаллических структур *E. coli*. **(А)** Гипотеза А: тРНК диффундирует быстро, маловероятно ее переиспользование. Ожидается отсутствие значимой автокорреляции. **(В)** Гипотеза В: тРНК диффундирует медленнее трансляции. Ожидается экспоненциальный спад автокорреляции со временем. **(С)** Гипотеза (С): тРНК остается связанной с рибосомой. Спад автокорреляции происходит медленнее, чем в модели В. Рисунок по статье [Cannarozzi et al., 2010].

#### 1.2.1.5. Связь между скоростью трансляции и точностью синтеза и фолдинга белка

Как было показано, использование предпочтительных кодонов в генах повышает эффективность элонгации трансляции. Однако, повышение скорости элонгации трансляции может привести к негативным эффектам, таким как снижение точности синтеза (более частые ошибки) или неправильный фолдинг белковой молекулы. Это может вызвать потерю белком функциональности. Поэтому кодонный состав генов подвергается давлению отбора, направленного не только на увеличение скорости элонгации трансляции, но и на правильность синтеза белка [Arava et al., 2003]. Так для *E. coli* и *B. subtilis* было установлено, что кластеры из редких кодонов преимущественно располагаются между участками, кодирующими домены



белков. Тем самым они способствуют снижению скорости трансляции в этих местах, чтобы обеспечить достаточно времени для правильного фолдинга синтезированного участка (Рисунок 7) [Zhang et al., 2009; Zhang and Ignatova, 2009; Siller et al., 2010].



**Рис. 7.** Предполагаемые места замедления трансляции находятся на границах между структурными доменами белков. Графики отображают скорость трансляции, цветные линии под ними – границы доменов в белках. Рисунок из статьи [Zhang and Ignatova, 2009].

Для определения влияния скорости элонгации и точности синтеза белка на кодонный состав, был разработан статистический метод, который позволяет сравнивать наборы последовательностей по НИК. Было проведено сравнение НИК в группах высоко- и низкоэкспрессирующихся генов, а также в группах консервативных и изменчивых сайтов в

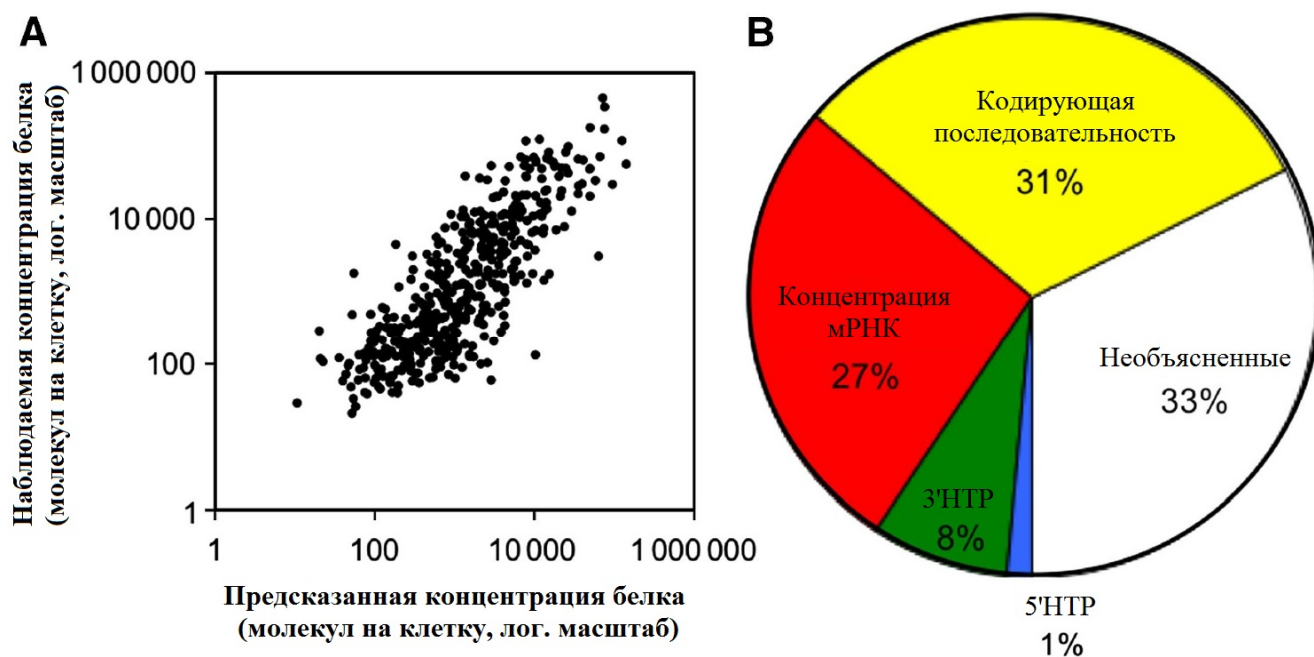
высокоэкспрессирующихся генах. В большой группе бактериальных геномов отчетливо прослеживается влияние как скорости, так и точности на НИК. Однако, влияние скорости, по-видимому, значительно сильнее, чем влияние точности, и является значимым в большей части геномов. Таким образом, отбор в сторону увеличения скорости трансляции является основным фактором, влияющим на неравномерность использования кодонов в быстро растущих бактериях, тогда как отбор в сторону увеличения точности играет небольшую дополнительную роль [Ran and Higgs, 2012].

#### **1.2.1.6. Кодонный состав генов и тканеспецифичная экспрессия у многоклеточных организмов**

Различные исследования одноклеточных и многоклеточных организмов показали, что кодонный состав играет значимую роль в эффективности трансляции посредством адаптации к пулу тРНК. Тем не менее, у человека и других млекопитающих роль кодонного состава все еще остается под вопросом ввиду наличия различных исследований с противоречивыми результатами [Waldman et al., 2010]. Анализ клеточного лизата клеток медуллобластомы Daou человека показал, что концентрация мРНК объясняет примерно 27% изменчивости концентрации белка, 30% объясняются различными особенностями кодирующей последовательности и около 10% объясняются вторичными структурами в 5' и 3' не транслируемых районах (Рисунок 8) [Vogel et al., 2010]. Вероятнее всего, тканеспецифичный синтез белков не позволяет обнаружить какие-либо общие особенности последовательностей, влияющие на эффективность трансляции.

В геноме человека аннотировано более 450 генов тРНК. Был разработан метод микрочипов для оценки количества тРНК, основанный на технике флуоресцентного мечения. При помощи данного метода было показано, что количество тРНК среди всей клеточной РНК значительно варьирует среди восьми различных типов тканей. Клетки мозга экспрессируют наибольший уровень тРНК, кодируемых в ядре, по сравнению с любой другой исследованной тканью, а также наибольший уровень митохондриальных тРНК. Обнаружены тканеспецифичные различия в экспрессии отдельных видов тРНК. Некоторые тРНК, декодирующие аминокислоты с похожими химическими свойствами, показали скоординированную экспрессию в отдельных типах тканей. Относительное количество тРНК и изоакцепторных тРНК статистически значимо коррелируют с кодонным составом высокоэкспрессируемых тканеспецифичных генов в группе тканей. Все эти результаты говорят о тканеспецифичной экспрессии тРНК, что хорошо подтверждает роль гетерогенности тРНК в регуляции трансляции [Dittmar et al., 2006].





**Рис. 8.** (А) Зависимость наблюдаемого количества белка от предсказанного.  $r = 0,67$  (логарифмический масштаб,  $P < 10^{-3}$ ). (В) Вклады различных особенностей в объяснение разнообразия количества белка. Рисунок по статье [Vogel et al., 2010].

При помощи индекса адаптации тРНК (tAI) [dos Reis et al., 2003] был проведен широкомасштабный тканеспецифичный анализ эффективности трансляции генов человека. В результате была обнаружена значимая корреляция между tAI и уровнем экспрессии генов как в отдельных тканях, так и в целом организме. Что интересно, обнаружена значительно более высокая корреляция во взрослых тканях в противоположность эмбриональным. Это говорит о том, что пул тРНК является более устойчивым во взрослом периоде. Дополнительный анализ показал, что коадаптация между пулом тРНК и кодонным составом происходит глобально (а не тканеспецифично). При использовании предполагаемых тканеспецифичных пулов тРНК было выявлено, что тканеспецифичные гены более адаптированы к ним, чем другие гены, и транслируются эффективнее. Аналогичные результаты получены для других млекопитающих [Waldman et al., 2010].

## 1.2.2. Влияние вторичных структур в мРНК на эффективность элонгации трансляции

### 1.2.2.1. Вторичные структуры в 5'-НТР и районе старт-кодона трансляции

Кроме частот кодонов на скорость движения рибосомы по мРНК могут влиять вторичные структуры (шпильки). Формирование данных структур в районе старт-кодона трансляции гена может приводить к снижению эффективности инициации трансляции или даже к ее полному выключению.

При исследовании гена *lamB* *E. coli*, являющегося белком внешней мембраны, были обнаружены две мутации: одна в позиции -2 относительно старта кодирующей части гена, другая – в +6. Данные мутации способствовали образованию шпильки в данном районе и блокированию последовательности Шайна-Дальгарно. Это приводило к недоступности последней для рибосомы и отсутствию инициации трансляции [Hall et al., 1982].

Для *E. coli* показано, что шпильки со свободной энергией -6 ккал/моль и выше не вызывают уменьшения эффективности трансляции. Однако, при образовании вторичной структуры с энергетической стабильностью ниже этого значения, эффективность экспрессии падает в 10 раз на каждые -1,4 ккал/моль [de Smit and van Duin, 1994].

Исследования, проводимые на генах эукариот, также показали влияние шпилек на трансляцию. Однако шпильки должны обладать достаточной стабильностью. При модификации нуклеотидной последовательности гена таким образом, чтобы ATG кодон был заключен в шпильку с энергией Гиббса  $\Delta G = -30$  ккал/моль, снижение уровня кодируемого белка не наблюдалось. Однако при повышении стабильности шпильки до  $\Delta G = -50$  ккал/моль происходило снижение количества синтезируемого белка [Kozak, 1986].

Для эукариотических мРНК были экспериментально показаны два типа внутримолекулярных вторичных структур, которые ингибируют трансляцию в гене *cis*. Первые предотвращают формирование инициаторного комплекса, вторые – останавливают его продвижение по мРНК. Одна петля ( $\Delta G = -30$  ккал/моль) препятствует движению 40S субъединицы рибосомы, когда образуется на расстоянии 12 нуклеотидов от кэпа. Однако такая же петля не оказывает ингибирующего влияния, располагаясь на расстоянии 52 нуклеотида от кэпа. Этот результат согласуется с ранее установленным фактом, что 40S субъединичный комплекс однажды связавшись с мРНК способен расплетать встречающиеся на пути вторичные структуры. Таким образом, трансляция наиболее чувствительна к наличию вторичных структур именно в сайте посадки рибосом, т.е. на 5'-конце мРНК [Kozak, 1989].

Петлевая структура ( $\Delta G = -61$  ккал/моль), расположенная на расстоянии 72 нуклеотида от кэпа оказывается слишком стабильной для того, чтобы быть расплетенной сканирующим 40S субъединичным комплексом рибосомы. Она не мешает посадке комплекса на мРНК, однако препятствует его дальнейшему продвижению. Контрольные эксперименты показали, что 80S субъединичный комплекс рибосомы во время элонгации способен расплести данную вторичную структуру [Kozak, 1989].

Несмотря на образование шпилек перед рибосомным комплексом, большинство из них не препятствует его дальнейшему продвижению. Это объясняется геликазной активностью рибосом, позволяющей им денатурировать шпильки длиной до 27 пар оснований с предсказанной температурой плавления  $T_m = 70^\circ\text{C}$ . Возможно, белки S3 и S4 (субъединицы

рибосомного комплекса) принимают участие в этом процессе, так как мутации в них снижают его эффективность [Takyar et al., 2005; Studer and Joseph, 2006].

Известно, что эффективность трансляции может регулироваться вторичными структурами в мРНК не напрямую, а через белки-посредники. В *Xenopus laevis* контроль трансляции мРНК, участвующих в гомеостазе железа, осуществляется белком IRP (iron regulatory protein), который связывается с IRE (internal ribosome entry) районом на мРНК. Данный район представляет собой небольшую петлю недалеко от кэпа. Когда в организме мало железа, белок IRP связан с IRE и не позволяет 40S субъединице рибосомы связаться с мРНК. Когда железа в организме достаточно, его ионы связываются с IRP, тем самым давая возможность 40S субъединице связаться с мРНК и инициировать трансляцию [Meijer and Thomas, 2002].

Обзор геномов 340 видов (бактерий, архей, грибов, растений, насекомых, рыб, птиц и млекопитающих) показал наличие снижения стабильности вторичных структур в районе старт-кодона трансляции почти для всех организмов. Уменьшение стабильности обратно пропорционально GC-составу генома. У прокариот уменьшение стабильности пропорционально оптимальной температуре роста организма. Внутри геномов есть изменчивость по энергии шпилек в районе старт-кодона в мРНК, которая коррелирует с GC-составом, неравномерностью кодонного состава и уровнем экспрессии гена. Однако для птиц и млекопитающих аналогичных трендов в полных геномах не обнаружено. Но в GC-богатых генах такая зависимость есть. Был сделан вывод, что сниженная стабильность вторичных структур в районе старт-кодона является универсальной особенностью клеточной жизни. Предполагается, что этот спад стабильности способствует эффективному распознаванию старт-кодона инициаторной тРНК [Gu et al., 2010].

#### 1.2.2.2. Вторичные структуры в кодирующей части мРНК

На скорость движения рибосомы по мРНК могут влиять вторичные структуры (шпильки), образующиеся перед рибосомным комплексом [Jacks et al., 1988; Dam et al., 1990; Thanaraj and Argos, 1996; Lopinski et al., 2000; Takyar et al., 2005; Tuller et al., 2011]. Поэтому, если в нуклеотидной последовательности гена встречается много локальных инвертированных повторов, которые потенциально могут образовать шпильки, то скорость трансляции такого гена может быть снижена.

У *H. pylori* не обнаружено корреляции между неравномерностью использования кодонов и эффективностью трансляции. Однако обнаружена достоверная корреляция ( $P > 95\%$ ) с количеством вторичных структур в мРНК [Vladimirov et al., 2007]. Также достоверная корреляция между количеством в клетке белка и количеством потенциальных вторичных

структур в кодирующей его мРНК была обнаружена у *Mycoplasma gallisepticum* (неопубликованные данные). Коэффициент корреляции варьирует от  $r = 0,19$  до  $r = 0,47$  с достоверностью в основном  $P > 99\%$  (но не ниже 90%) для разных фаз роста организма (от начала логарифмической фазы до стационарной фазы роста).

Кроме эффективности трансляции шпильки в мРНК могут также влиять на другие особенности данного процесса, например, на правильность считывания кодонов. Так у вируса RSV (Rous sarcoma virus) шпилька в мРНК вызывает сдвиг рибосомы на 1 нуклеотид в 5' направлении, что приводит к соответствующему сдвигу рамки считывания. Это необходимо для правильного декодирования аминокислотной последовательности белка *Gag-pol*, предшественника фермента, ответственного за обратную транскрипцию и интеграцию [Jacks et al., 1988].

Известно, что вторичные структуры в мРНК могут влиять на правильность сворачивания (фолдинга) синтезируемого белка [Thanaraj and Argos, 1996]. Так у *Nephila clavipes* показана возможная связь стабильности вторичных структур в мРНК, кодирующей белок Spidroin 1, со вторичной структурой последнего. Данный белок состоит из двух доменов. Первый образует  $\beta$ -складки, второй – аморфный. Считается, что более стабильные шпильки в начале мРНК способствуют замедлению трансляции и тем самым позволяют полипептиду правильно свернуться, образовав  $\beta$ -складку [Zama, 1999].

Было проведено исследование 107 мРНК *E. coli* и 125 мРНК человека с совпадающими вторичными структурами белка. Свободная энергия сегментов мРНК из разных вторичных структур белка значительно различалась. Энергия была ниже для участков мРНК, кодирующих правильные вторичные структуры ( $\alpha$ -спирали и  $\beta$ -складки), чем для других. Это говорит о том, что выбор кодонов в исходной мРНК, кодирующей регулярную структуру белка, способствует ее большей стабильности [Jia and Luo, 2006].

Компьютерный анализ в дрожжах с кодонным разрешением показал, что вторичные структуры в мРНК замедляют процесс элонгации трансляции, чтобы обеспечить правильность синтеза белка. Позиционное специфическое влияние фолдинга мРНК на точность трансляции показано экспериментально при помощи замены синонимичных кодонов в трансгенах дрожжей [Yang et al., 2014].

Для некоторых организмов показано отсутствие влияния вторичных структур на эффективность трансляции. При анализе двух наборов микрочиповых данных по *D. melanogaster* и *E. coli* при помощи множественной линейной регрессии с использованием линейной модели было установлено, что неравномерность частот кодонов и GC-состав 1-ой, 2-ой и 3-ей позиций нуклеотидов значительно влияет на эффективность экспрессии генов. Значимого влияния стабильности вторичных структур в мРНК обнаружено не было [Eck and

Stephan, 2008]. Как видно из раздела 1.2.1. главы «Обзор литературы» и как будет показано в главе «Результаты и обсуждение», в *E. coli* определяющим фактором эффективности элонгации трансляции являются именно частоты кодонов, а не вторичные структуры. Аналогичные результаты относятся и к *D. melanogaster* (в этой работе данные не приводятся). Возможно, именно поэтому исследователям не удалось обнаружить указанной зависимости.

Шпильки могут принимать участие в защите мРНК от разрушения нуклеазами, уменьшая их активность [Diwa et al., 2000]. Защитой также может быть обеспечена и ДНК. Скорость мутаций в одноцепочечной ДНК выше, чем в двуцепочечной, поэтому образование шпилек в ней снижает вероятность возникновения мутаций. Отбор на шпильки идет сильнее в высокоэкспрессирующихся генах, так как их ДНК дольше находится в одноцепочечном состоянии [Hoede et al., 2006].

Считается, что вторичные структуры в открытых рамках считывания снижают эффективность трансляции, и организмы должны стараться их избегать [Klionsky et al., 1986; Guisez et al., 1993; Schmittgen et al., 1994]. Однако есть организмы, у которых кодирующие последовательности содержат стабильные вторичные структуры, например, *Bacillus subtilis* [Rocha et al., 1999].

При исследовании 51 последовательности мРНК генов различных организмов (прокариот, растений, беспозвоночных и высших животных) было установлено, что свободная энергия исходной мРНК ниже, чем у мРНК с теми же составом нуклеотидов и длиной, но случайным порядком нуклеотидов (рандомизированной). Для рандомизации последовательности использовалась программа SHUFFLE от UWGCG. Данная программа случайным образом переставляет нуклеотиды в последовательности, оставляя их состав постоянным. Результат говорит о том, что при выборе кодонов предпочитают те, которые способствуют большей стабильности мРНК [Seffens and Digby, 1999].

Позднее это исследование было перепроверено другой группой ученых на этом же наборе данных. Исследователи сделали вывод, что наблюдаемая более высокая стабильность мРНК структур в исходных последовательностях может быть объяснена различием в динуклеотидном составе исходной и рандомизированной мРНК, так как термодинамика фолдинга мРНК в сильной степени зависит от динуклеотидных взаимодействий [Workman and Krogh, 1999].

Чтобы окончательно выяснить, идет ли отбор кодонов в генах в сторону оптимизации вторичной структуры мРНК, третьей группой исследователей был разработан новый алгоритм рандомизации первичной последовательности гена (DicodonShuffle). Программа позволяет получить из исходной новую последовательность со случайным порядком нуклеотидов, но с сохранением кодируемой аминокислотной последовательности, частот кодонов и частот

динуклеотидов. В результате было показано, что для генов 10-ти из 14-ти исследованных видов эубактерий и 1 эукариота (*S. cerevisiae*) есть статистически значимое отклонение в сторону предпочтения вторичных структур в мРНК. Эти результаты были интерпретированы с точки зрения возможной роли вторичных структур в процессинге РНК, регуляции стабильности мРНК и регуляции трансляции [Katz and Burge, 2003].

Новые экспериментальные методы позволяют более подробно изучать влияние вторичных структур в мРНК на скорость элонгации трансляции. Такие структуры используются клеткой для создания разнообразных стратегий регуляции трансляции, таких как программируемый сдвиг рамки считывания, регуляция уровня экспрессии белков, локализация рибосомы и ко-трансляционный фолдинг белка. Анализ одиночных молекул при помощи оптического пинцета при изучении шпилек в мРНК показал, что скорость трансляции идентичных кодонов в декодирующем центре находится под большим влиянием GC-состава вторичных структур в месте входа мРНК в рибосому [Qu et al., 2011].

Количественный анализ геликазной активности показал, что рибосома, в отличие от ранее исследованных геликаз, использует два различных активных механизма расплетания вторичных структур в мРНК. Первый механизм заключается в предпочтительном взаимодействии рибосомы с участком мРНК, содержащем шпильку в расплетенном (открытом) состоянии. Т.е. во время температурных флуктуаций состояния шпильки между закрытой и открытой рибосома чаще взаимодействует с открытым состоянием. Для других геликаз было показано, что такой механизм не требует внешнего источника энергии. Второй механизм заключается в механическом расплетании шпильки под воздействием конформационных изменений, которые сопровождают транслокацию рибосомы. Т.е. в данном случае рибосома прикладывает определенное усилие и, соответственно, затрачивает энергию. Данный механизм обеспечивает минимальную базальную скорость трансляции в клетке [Qu et al., 2011].

В работе [Xie, 2013] показано, что для объяснения имеющихся экспериментальных данных по эффективности трансляции достаточно только второго механизма (механического разделения цепей) расплетания шпилек перед рибосомой из двух, предложенных в статье [Qu et al., 2011]. На основе созданной модели были предсказаны скорости трансляции нескольких мРНК. Экспериментальные подтверждения этих предсказаний позволили бы более точно ответить на вопрос о механизме геликазной активности рибосомы [Xie, 2013].

### **1.2.2.3. Современные экспериментальные методы определения вторичных структур в мРНК**

Компьютерные программы для расчета вторичных структур мРНК позволяют анализировать большое количество нуклеотидных последовательностей, вплоть до целых

геномов. Однако точность результатов такого анализа пока ниже, чем у экспериментальных данных. До недавнего времени определение вторичной структуры транскрипта даже самого маленького организма представляло собой большую задачу для экспериментаторов. Но теперь разработан новый экспериментальный метод PARS (parallel analysis of RNA structure), позволяющий определить вторичную структуру тысяч молекул РНК с разрешением в один нуклеотид. При помощи данного метода получены профили для 3000 различных транскриптов *S. cerevisiae*. Анализ этих данных выявил несколько особенностей РНК: повышенное содержание вторичных структур в кодирующей части, по сравнению с нетранслируемыми флангами; периодичность участия нуклеотидов в формировании вторичных структур с периодом 3 нуклеотида; отрицательная корреляция между эффективностью трансляции мРНК и вторичной структурой в районе старта трансляции [Kertesz et al., 2010].

Недавно был разработан новый метод полногеномного исследования вторичных структур РНК *in vivo* с разрешением в один нуклеотид. С его помощью было проведено исследование более чем 10000 транскриптов *Arabidopsis thaliana*. Результаты показали наличие периодичности с периодом 3 нуклеотида в структуре кодирующих районов (как уже было показано для *S. cerevisiae* [Kertesz et al., 2010]), наличие вторичных структур с высокой и низкой энергией в сайтах альтернативного полиаденилирования, а также наличие стабильных вторичных структур в 5' сайтах сплайсинга, которые коррелируют с отсутствием сплайсинга [Ding et al., 2014].

### **1.3. Биоинформатические методы оценки эффективности экспрессии генов**

#### **1.3.1. Математические индексы**

Математический подход к изучению эффективности экспрессии генов основывается на создании математических моделей данного процесса. Модели описываются формулами, содержащими в своем составе в качестве параметров различные характеристики нуклеотидных последовательностей (например, частоты кодонов или дикодонов, длину гена, количество повторов и т.д.). Такие формулы-модели называют индексами, т.к. в результате их расчетов получается некоторое число, индекс, характеризующее эффективность исследуемого процесса.

Одним из самых успешно применяемых индексов является индекс адаптации кодонов CAI (codon adaptation index). Он рассчитывается для каждого гена на основе частот входящих в него кодонов (см. Формулу 4). Данный индекс позволяет предсказывать эффективность экспрессии генов, оценивать уровень адаптации генов вирусов к их хозяевам, а также сравнивать использование синонимичных кодонов в разных организмах. Дополнительно CAI позволяет оценивать эффективность экспрессии гетерологичных генов [Sharp and Li, 1987].

Индекс CAI не является первым, и до него уже было предложено несколько величин, позволяющих оценивать неравномерность использования кодонов. Одной из таких мер был предложенный McLachlan с коллегами индекс неравномерности предпочитаемых кодонов (codon preference bias), позволяющий оценивать степень отклонения наблюдаемых частот от некоторого заранее заданного паттерна использования [McLachlan et al., 1984]. Другой подход заключался в расчете величины  $\chi^2$  для оценки отклонения от равномерного использования кодонов и последующем нормировании ее на длину гена. Этот метод позволял сравнивать друг с другом различные гены [Sharp et al., 1986].

Еще один подход состоял в оценке относительной пользы различных кодонов с точки зрения эффективности трансляции. Например, Ikemura обнаружил «оптимальные» кодоны в *E. coli* и *S. cerevisiae*, которые, как предполагалось, транслируются более эффективно, чем другие, и рассчитал их частоты в генах данных организмов [Ikemura, 1981, 1982, 1985]. Практически тот же смысл имеет индекс CBI (codon bias index – индекс смещения кодонов), предложенный Bennetzen и Hall для исследования генов дрожжей [Bennetzen and Hall, 1982]. Данные индексы в определенной степени полезны, но имеют ряд недостатков. Во-первых, некоторые аминокислоты обычно исключаются из анализа из-за того, что не понятно, какой кодон является оптимальным. Во-вторых, все кодоны разделены только на две категории: оптимальные и неоптимальные. Это не позволяет сравнивать кодоны из одной категории. В-третьих, нет хорошего базиса для сравнения кодонов между организмами [Sharp and Li, 1987].

Еще один индекс – статистика предпочтения кодонов (codon preference statistic) [Gribbskov, 1984]. Эта статистика основана на отношении вероятности обнаружить определенный кодон в высокоэкспрессирующемся гене к вероятности обнаружить его в случайной последовательности с таким же нуклеотидным составом. Было показано, что данная статистика может использоваться для локализации генов в последовательности ДНК, для предсказания относительного уровня их экспрессии и для поиска ошибок секвенирования. Однако она не нормализована и поэтому не может быть применена для сравнения двух последовательностей.

CAI индекс похож на статистику предпочтения кодонов, однако, в отличие от нее, является нормализованным и позволяет сравнивать последовательности как внутри одного генома, так и между разными геномами. Однако данный индекс, хоть и рассчитывается по частотам кодонов в генах, требует наличия предварительной (опорной) выборки генов, для которых установлено, что они являются высокоэкспрессирующимися.

$$CAI = \left( \prod_{k=1}^L w_k \right)^{1/L}, \quad (4)$$



$$w_k = \frac{x_{i,j}}{y_j}, \quad (5)$$

где  $x_{i,j}$  – число попаданий кодона  $i$ , кодирующего аминокислоту  $j$ , в опорную выборку;  $y_j$  – число попаданий самого оптимального кодона для этой аминокислоты в опорную выборку;  $L$  – количество кодонов в опорной выборке.

ENc – эффективное число кодонов (effective number of codons, или ENc), используемых в гене [Wright, 1990]. Данный индекс оценивает число используемых кодонов, которые сформировали бы такую же неравномерность кодонного состава, как наблюдаемая. ENc может быть вычислен без использования заданного набора оптимальных кодонов. Это общая мера неравномерности безотносительно к равновероятному использованию синонимичных кодонов.

$$ENc = 2 + (9/F^2) + (1/F^3) + (5/F^4) + (3/F^6), \quad (6)$$

$$F = \frac{n}{n-1} \sum p_i^2, \quad (7)$$

где  $n$  – число кодонов, использованных в гене для данной аминокислоты, а  $p_i$  – частота использования  $i$ -го кодона (относительно синонимичных ему),  $p_i = n_i/n$ . В гене с большой неравномерностью кодонного состава, где используется только один кодон для каждой аминокислоты, этот индекс равен 20. В гене с равновероятным использованием кодонов ENc = 61.

Индекс RSCU (relative synonymous codon usage) [Sharp et al., 1986] рассчитывается как отношение наблюдаемой частоты кодона к частоте, ожидаемой при условии равновероятного использования кодонов в пределах синонимичной группы:

$$RSCU_i = \frac{X_i}{\frac{1}{n_i} \sum_{j=1}^{n_i} X_j}, \quad (8)$$

где  $X_i$  – число встречаемости  $i$ -кодона в гене,  $n_i$  – число альтернативных синонимичных кодонов для данной аминокислоты (от 2 до 6). Значения RSCU близкие к 1,0 означают отсутствие неравномерности использования кодона. Значения больше 1,0 соответствуют повышенной частоте использования кодона, меньше 1,0 – соответствуют пониженной частоте. К достоинствам индекса RSCU относится его независимость от аминокислотного состава гена, что позволяет использовать его для сравнения генов или групп генов, различающихся размером и аминокислотным составом.

Индекс CBI является мерой направленного смещения кодонного состава в сторону набора оптимальных кодонов [Bennetzen and Hall, 1982]:

$$CBI = \frac{N_{opt} - N_{ran}}{N_{tot} - N_{ran}}, \quad (9)$$

где  $N_{opt}$  – число оптимальных кодонов,  $N_{tot}$  – число синонимичных кодонов,  $N_{ran}$  – ожидаемое число оптимальных кодонов, если бы кодоны использовались равновероятно. Индекс CBI близок к индексу  $F_{op}$  (см. далее) за тем исключением, что  $N_{ran}$  используется как шкалирующий фактор. В гене с очень высокой неравномерностью кодонного состава CBI может достигать значения 1,0. При равномерном использовании кодонов в гене его значение CBI равно 0.

Индекс  $F_{op}$  (frequency of optimum codons) оценивает частоту оптимальных кодонов в гене [Ikemura, 1981; Ikemura, 1985; Ikemura and Ozeki, 1982]. Это видоспецифичная мера неравномерности кодонного состава, представляющая собой отношение частоты оптимальных кодонов к общему числу синонимичных кодонов. Его значение варьирует от 0 (ген не содержит оптимальных кодонов) до 1,0 (ген полностью состоит из оптимальных кодонов).

Компьютерное сравнение нескольких индексов ( $N_c$  – effective number of codons, CBI – codon bias index, CAI – codon adaptation index, “Scaled”  $\chi^2$ , ICBI – intrinsic codon bias index) показало, что при сравнении кодирующих последовательностей с разной длиной только  $N_c$  и CAI дают эквивалентные значения по всем длинам кодирующих последовательностей и частотам использования нуклеотидов в третьей позиции [Comeron and Aguadé, 1998].

Одним из самых часто применяемых и точных индексов адаптации кодонов на данный момент является индекс адаптации тРНК tAI (tRNA adaptation index) [dos Reis et al., 2003; dos Reis et al., 2004; Waldman et al., 2010]. Данный индекс основан на ко-адаптации кодонного состава высокоэкспрессируемых генов и пула тРНК. А также на корреляции между количеством определенных тРНК в клетке и количеством копий их генов в геноме организма. Данный индекс был впервые предложен в статье [dos Reis et al., 2003]. Для его расчета сначала для каждого кодона требуется вычислить специальную величину, называемую абсолютной адаптивностью:

$$W_i = \sum_{j=1}^n (1 - s_{ij}) \cdot tGCN_{ij}. \quad (10)$$

Здесь  $tGCN_{ij}$  – это количество копий гена соответствующей изоакцепторной тРНК для  $i$ -го кодона, а  $s_{ij}$  – селективные ограничения на эффективность кодон-антикодонных взаимодействий для  $j$ -ой тРНК. Для естественных изоакцепторов  $s = 0$ , для изоакцепторов с неоднозначными парами оснований  $s = 0,5$ . После расчета всех значений, относительная адаптивность кодона вычисляется по формуле:

$$w_i = W_i / W_{max}, \quad (11)$$

где  $W_{max}$  – наибольшее значение адаптивности. В итоге индекс адаптации тРНК для последовательности рассчитывается как геометрическое среднее  $w_i$  для каждого кодона в последовательности:

$$tAI = (\prod_{k=1}^L w_k)^{1/L}, \quad (12)$$

где  $L$  – длина в кодонах кодирующей последовательности (за исключением стоп-кодона) [dos Reis et al., 2003].

Неравномерное использование кодонов формируется благодаря совместному влиянию трансляционного отбора и мутаций (мутационного смещения). Современные методы для анализа НИК применяют эвристики для учета обоих компонентов. Это ограничивает возможности анализа НИК, как метода изучения эволюции и оптимизации нуклеотидных последовательностей белок-кодирующих генов. Поэтому был разработан новый самосогласованный индекс НИК scnRCA, который включает неявную коррекцию мутационного смещения, облегчая исследование влияния только трансляционного отбора на НИК [O'Neill et al., 2013].

Данный индекс рассчитывается по формуле:

$$nRCA(G) = (\prod_{c=1}^L \omega_n(G_c))^{1/L}, \quad (13)$$

$$\omega_n(G_c = xyz) = \frac{\omega(xyz)}{\max_{ijk \rightarrow G_{[c]}}(\omega(ijk))}, \quad (14)$$

$$\omega(xyz) = \frac{f(x,y,z)}{f_1(x)f_2(y)f_3(z)}, \quad (15)$$

где  $f(x,y,z)$  – наблюдаемая частота кодона,  $f_1(x)f_2(y)f_3(z)$  – ожидаемая частота кодона [O'Neill et al., 2013]. Расчет scnRCA производится на основании значений nRCA итеративно с пошаговым уменьшением опорной выборки до заранее заданного размера и далее до стабилизации ее состава генов [Carbone et al., 2005].

Индекс scnRCA был проверен на экспериментальных данных по экспрессии генов и применен для детального анализа НИК в *Pseudomonadales*. Результаты показали, как происходит отбор специфических кодонов в высокоэкспрессирующихся генах в контексте полногеномного смещения частот кодонов, и как баланс между мутационным и трансляционным смещениями приводит к различным определениям оптимальности кодона [O'Neill et al., 2013].

Анализ других умеренно и быстро растущих бактерий подтверждает гипотезу, что в высокоэкспрессируемых генах у аминокислот, кодируемых двумя кодонами, кодоны, имеющие в третьей позиции нуклеотиды С и А, встречаются чаще других. Аналогично у аминокислот, кодируемых четырьмя кодонами, кодоны, имеющие в третьей позиции нуклеотид U, встречаются наиболее часто. Использование несмещенного индекса scnRCA для оценки НИК позволило впервые показать, что особенности трансляционного отбора очень консервативны у

*Pseudomonadales* в свете значительных изменений в составе генома и распространяются далеко за основной набор высоко оптимизированных генов в каждом геноме [O'Neill et al., 2013].

Используемый в данной работе индекс эффективности элонгации EEI (elongation efficiency index) был разработан в ИЦиГ СО РАН В. А. Лихошваем и Ю. Г. Матушкиным [Лихошвай и Матушкин, 2000; Likhoshvai and Matushkin, 2002]. Его расчет основан на учете кодонного состава генов и количестве или энергетической стабильности вторичных структур в мРНК. Более подробное описание см. в разделе 2.2. главы «Методы и алгоритмы». Отличительной особенностью данного индекса является отсутствие необходимости в дополнительных исследованиях для получения набора высокоэкспрессирующихся генов, а также учет потенциальных вторичных структур в мРНК. Последний факт позволяет значительно расширить область применимости данного индекса по сравнению с другими индексами, которые учитывают только кодонный состав генов [Владимиров и др., 2007].

### **1.3.2. Программы для анализа контекстных характеристик нуклеотидных последовательностей**

Современное программное обеспечение намного упрощает работу исследователей. Оно не только дает возможность работать с невероятно большими объемами биологических данных, но и позволяет предсказывать или оценивать некоторые параметры, которые пока не имеют экспериментально определенных значений. Например, сейчас существуют программы, позволяющие абсолютно для любой исходной нуклеотидной последовательности предсказать ее вторичную структуру, или оценить уровень экспрессии гена по частотам используемых в нем кодонов.

#### **1.3.2.1. Программы для анализа кодонного состав нуклеотидных последовательностей**

**CodonW** – программа, созданная John Peden для упрощения многопараметрического анализа использования кодонов и аминокислот. Она позволяет рассчитывать основные индексы использования кодонов (CAI,  $F_{op}$ ,  $N_c$ , CBI), аминокислотные индексы (GRAVY score, Aromaticity), производить анализ соответствия по таким параметрам, как использование кодонов, RSCU, использование аминокислот. Программа доступна для скачивания по адресу: <http://codonw.sourceforge.net/>.

Следующий набор программ служит уже больше не цели расчета параметров последовательности, а конструирования самой последовательности.

**Codon optimizer** – программа, которая позволяет оптимизировать кодонный состав исходной последовательности под требуемые частоты использования кодонов. Например, при

переносе чужеродного гена в организм нового хозяина, можно оптимизировать кодонный состав гена под предпочтительные кодоны у трансформируемого организма. Предпочтительные кодоны хозяина определяются на основе выборки из его высокоэкспрессирующихся генов. Дополнительно программа также позволяет рассчитывать индекс CAI и идентифицировать внутренние сайты рестрикции II типа в оптимизированной последовательности [Fuglsang, 2003].

**JCat** – программа, похожая по функциональности на предыдущую. Основной функцией программы является оптимизация кодонного состава целевого гена для его экспрессии в организме-хозяине. Также программа позволяет исключить из последовательности различные сайты рестрикции и rho-независимые терминаторы транскрипции. Дополнительно программа позволяет рассчитать индекс CAI. Основное отличие JCat от других аналогичных программ состоит в том, что она не требует дополнительной информации о наборе высокоэкспрессирующихся генов и, таким образом, является быстрым и простым инструментом. Программа доступна по адресу: <http://www.jcat.de> [Grote et al., 2005].

**GeneDesign** – программное обеспечение, представляющее собой набор модулей и способствующее быстрому дизайну искусственных генов многокилобазной длины. Это является особенностью данной программы и позволяет конструировать целые хромосомы с необходимыми исследователю параметрами, например, кодонным составом. Адрес приложения в интернете: <http://slam.bs.jhmi.edu/gd> [Richardson et al., 2006].

**Gene Designer** – это приложение для быстрого и простого создания синтетических ДНК сегментов. Пользователь может легко добавлять, редактировать и комбинировать генетические элементы (промоторы, открытые рамки считывания, тэги) посредством интуитивного «drag-and-drop» интерфейса. Использование преимущества оптимизированного алгоритма позволяет открытым рамкам считывания внутри ДНК конструкции быть оптимизированными для эффективной экспрессии в любом организме-хозяине [Villalobos et al., 2006].

**OPTIMIZER** – on-line приложение, которое оптимизирует кодонный состав гена для увеличения эффективности его экспрессии. Доступно три метода оптимизации: «одна аминокислота – один кодон»; метод, основанный на алгоритме Монте-Карло; а также метод, производящий максимальную оптимизацию с минимальными изменениями в исходной последовательности. Особенностью данной программы является возможность проведения оптимизации ДНК последовательности с использованием заранее составленных таблиц частот кодонов для более чем 150 прокариотических видов в условиях сильного трансляционного отбора. Также программа позволяет избегать заданных сайтов рестрикции. Сайт программы: <http://genomes.urv.es/OPTIMIZER> [Puigbò et al., 2007].

**GeneGenie** является одной из самых новых программ, позволяющих оптимизировать первичную структуру моделируемых генов и генетических конструкций для увеличения эффективности их экспрессии. Основным отличием данной программы от других является возможность создания библиотек вариантов синтетических генов и конструкций для проведения эволюционных исследований. Программа доступна по адресу: <http://www.gene-genie.org> [Swainston et al., 2014].

### 1.3.2.2. Программы для анализа вторичных структур в нуклеотидных последовательностях

Понятие вторичная структура РНК впервые появилось в работах Doty и Fresco [Doty et al., 1959; Fresco et al., 1960]. Предсказание вторичной структуры посредством минимизации энергии с использованием метода ближайшего соседа началось с работ Tinoco и его коллег [Tinoco et al., 1971, 1973; Uhlenbeck et al., 1973; Borer et al., 1974], а также с Delisi и Crothers [Delisi and Crothers, 1971]. Эффективные алгоритмы предсказания с применением динамического программирования были разработаны независимо разными исследователями [Waterman, 1978; Waterman and Smith, 1978; Nussinov et al., 1978; Nussinov and Jacobson, 1980; Zuker and Stiegler, 1981; Sankoff et al., 1983].

**Vienna RNA** – сервер, предоставляющий веб-интерфейс для доступа к наиболее востребованным функциям одноименной программы. Данный веб-сервис позволяет предсказывать вторичную структуру заданной последовательности нуклеотидов и консенсусную вторичную структуру набора из нескольких последовательностей. Кроме этого у данного сервиса есть полезная экспериментаторам функция, позволяющая конструировать последовательность с заданной наперед структурой. Веб-сервер Vienna RNA доступен по ссылке <http://rna.tbi.univie.ac.at/> [Hofacker, 2003].

**Mfold** – веб-сервер, обеспечивает доступ к нескольким связанным приложениям, позволяющим предсказывать вторичную структуру одноцепочечных нуклеотидных последовательностей. Аналоги данного ресурса выдают в результате меньше информации, однако за меньшее время и для сотен последовательностей одновременно. Адрес веб-сервера: <http://www.bioinfo.rpi.edu/applications/mfold> [Zuker, 2003].

**SeqFold** – совмещает высокопроизводительную обработку экспериментальных данных по вторичным структурам РНК и проведение полногеномного предсказания вторичных структур РНК. Программа преобразовывает экспериментально определенные вторичные структуры РНК в предустановленные профили и использует последние для реконструкции вторичных структур в исследуемых РНК. SeqFold способен обрабатывать данные из различных экспериментов (PARS, SHAPE-Seq, FragSeq, SHAPE). Результаты применения программы

SeqFold для реконструкции вторичных структур транскриптома дрожжей показало различное влияние вторичных структур на регуляцию экспрессии генов (транскрипцию, трансляцию и взаимодействие белков с мРНК) [Ouyang et al., 2013].

**mRNA-optimiser** – совсем недавно разработанное приложение, которое позволяет оптимизировать имеющуюся мРНК для увеличения (или уменьшения) свободной энергии ее вторичных структур без изменения кодируемой аминокислотной последовательности. Алгоритм характеризуется низкой затратой времени за счет применения упрощенного приближения формирования петель. Данные показывают, что метод позволяет увеличить минимум свободной энергии в среднем более чем на 40%, уменьшив тем самым стабильность шпилек. Сайт приложения: <http://bioinformatics.ua.pt/software/mRNA-optimiser/> [Gaspar et al., 2013].

### **1.3.3. Изучение процесса трансляции при помощи математических стохастических моделей**

Одним из методов изучения процесса трансляции является создание его теоретических моделей, основанных на различных экспериментальных данных. Современные компьютеры позволяют моделировать трансляцию и получать в результате такие данные как, например, скорость/эффективность трансляции или плотность (среднее число на единицу длины) рибосом вдоль мРНК.

Была разработана теоретическая модель синтеза белка на рибосоме с использованием искусственно созданного шаблона мРНК, который содержит гомогенную последовательность (т.е. все кодоны которой идентичны). В рамках этой модели принято, что в окружающей среде присутствуют только две аминокислоты, причем только одна из них соответствует аминокислоте, закодированной в данной мРНК. Поскольку стохастическое движение рибосомы вдоль мРНК состоит из пауз и транслокаций, под временем жизни рибосомы на кодоне подразумевается сумма продолжительностей паузы и последующей транслокации на данном кодоне. Рибосома обладает системой контроля качества, которая проверяет правильность декодирования аминокислоты [Sharma and Chowdhury, 2011].

При помощи данной модели было исследовано влияние системы контроля качества на распределение времени жизни рибосомы на кодонах в мРНК. Также было изучено влияние остановки движения рибосом (заторов), вызванных их стерическим взаимодействием, на данное распределение. В результате было получено аналитическое выражение для распределения времен жизни рибосом в модели. Показано, что чем выше плотность рибосом (количество штук на единицу длины мРНК), тем больше время жизни рибосомы на кодоне. Теоретические результаты согласуются с доступными экспериментальными данными. Для проверки новых

оценок модели требуется проведение дополнительных экспериментов [Sharma and Chowdhury, 2011].

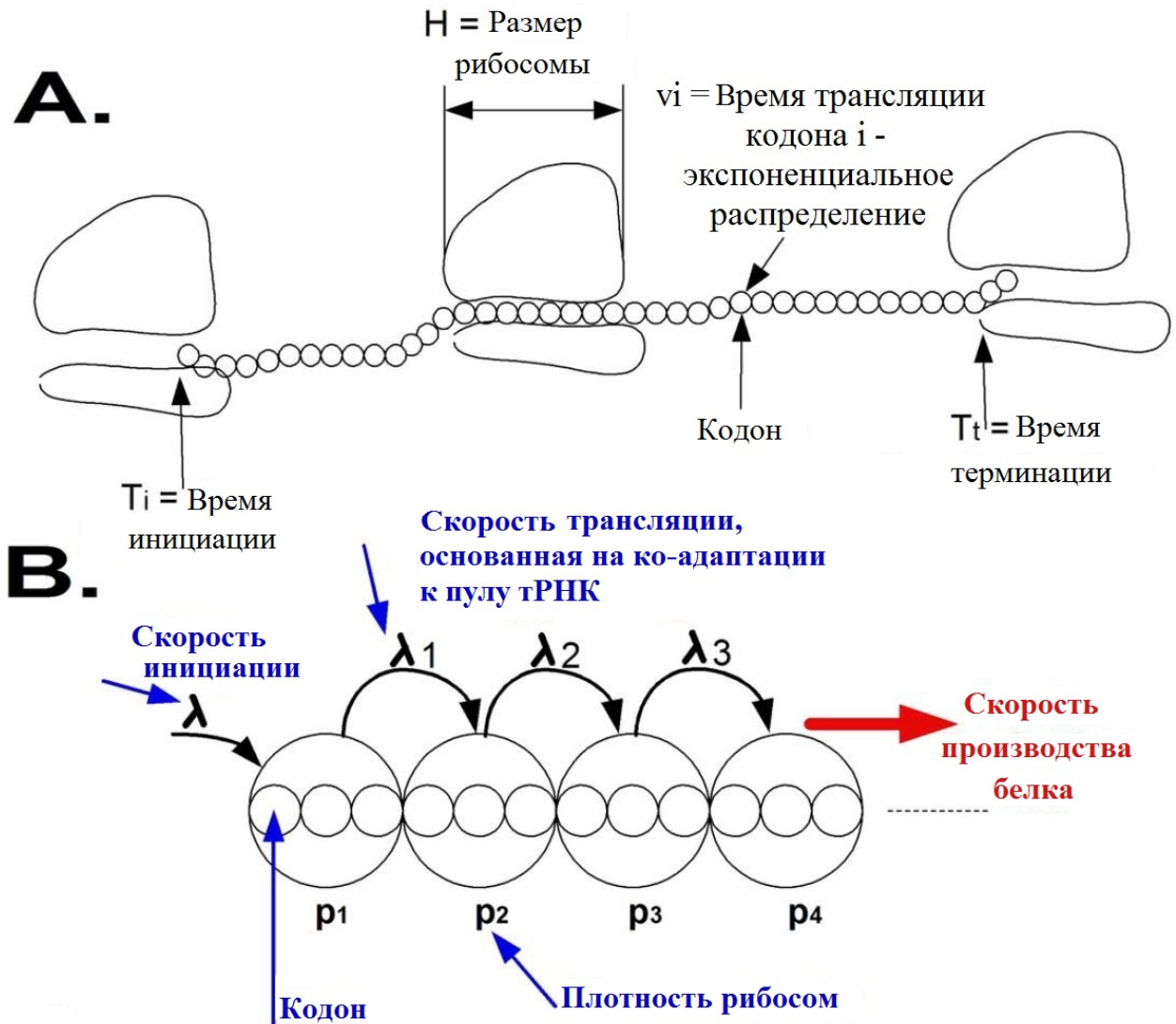
Как было отмечено ранее, рибосома обладает внутренней геликазной активностью, что обеспечивает трансляцию даже через достаточно стабильные шпильки, и позволяет устранять многие структуры, ограничивающие скорость трансляции. Таким образом, при движении по мРНК рибосомы расплетают вторичные структуры, и в мРНК может не оказаться шпилек, предсказанных алгоритмами фолдинга мРНК. Истинные структуры могут быть такими, которые способны сформироваться в свободных участках нуклеотидной последовательности (окнах) между рибосомами. Длина и время существования таких окон должны зависеть от кинетики трансляции и, возможно, значительно варьировать вдоль мРНК [Welch et al., 2009a].

Была разработана еще одна модель трансляции, которая учитывает физическую и динамическую природу данного процесса. Модель потока рибосом (МПР/RFM – ribosomal flow model) оценивает фундаментальные особенности процесса трансляции, включая скорости трансляции, уровни количества белков, плотности рибосом и взаимодействие всех этих параметров, лучше, чем альтернативные (не физические) методы (Рисунок 9). При помощи данной модели был проведен первый крупномасштабный анализ трансляции генов [Reuveni et al., 2011].

МПР может быть использована для точного расчета скорости инициации и энергетических затрат на трансляцию. Данные параметры не могли быть получены при помощи других более ранних моделей. Было обнаружено, что увеличение количества доступных рибосом (или, что эквивалентно, скорости инициации) увеличивает скорость геномной трансляции и среднюю плотность рибосом до определенного предела, после которого наступает насыщение. То есть дальше скорость практически не растет. Удивительно, но предположение о том, что система трансляции работает как раз при таких условиях, при которых почти наступает насыщение системы, позволяет МПР точнее описывать экспериментальные данные. Этот результат предполагает, что во всех организмах, которые были проанализированы (от бактерий до человека), глобальная скорость инициации близка к скорости полного насыщения. Тот факт, что аналогичные результаты не наблюдаются для гетерологичных генов, говорит о том, что эта особенность находится под влиянием отбора [Reuveni et al., 2011].

Примечательно, что производительность МПР и альтернативных моделей значительно различаются в случае гетерологичных генов. Это свидетельствует о перспективной биотехнологической значимости МПР в предсказании количества гетерологичных белков перед их непосредственной экспрессией в целевом организме [Reuveni et al., 2011].



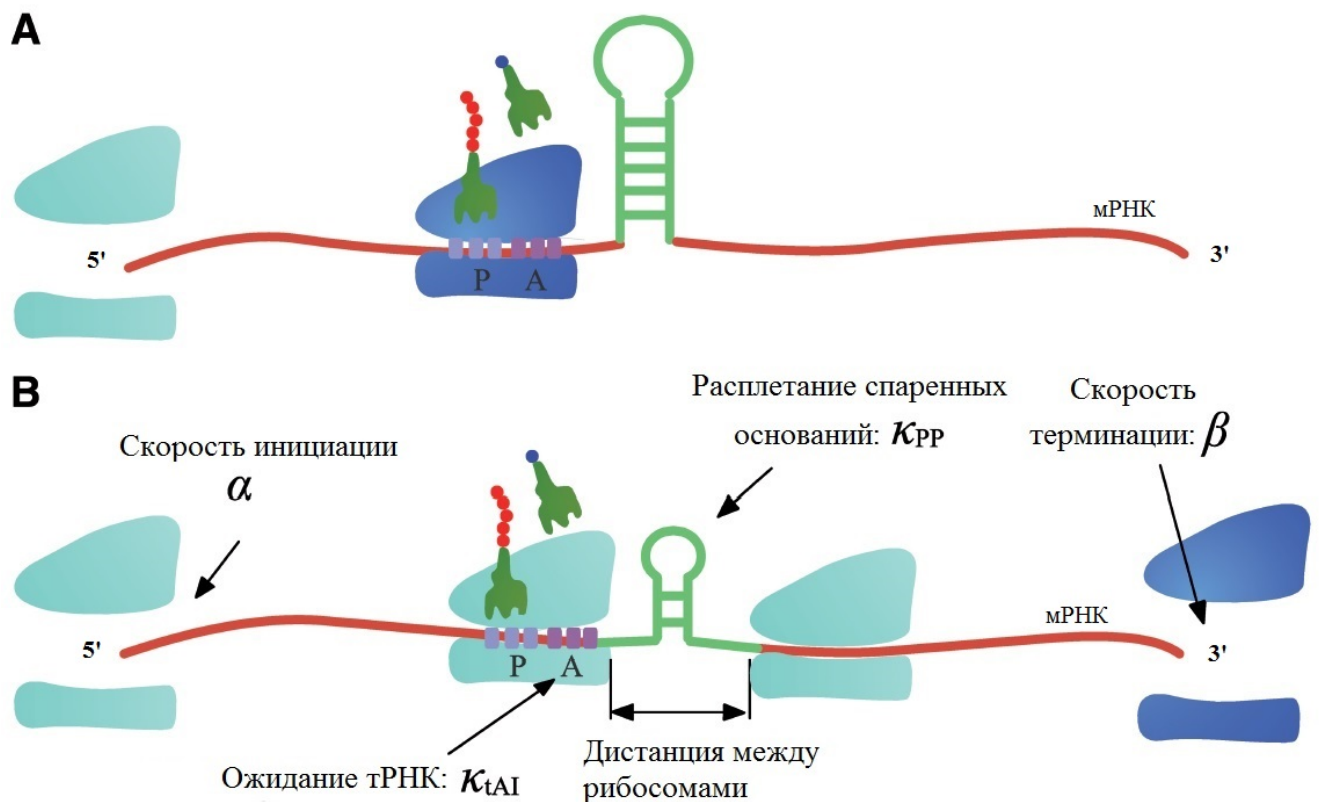


**Рис. 9.** Основные параметры модели потока рибосом (МПП). (А) Модель: каждый кодон характеризуется экспоненциально распределенным временем трансляции, рибосомы имеют размер и могут блокировать друг друга. (В) МПП имеет два свободных параметра: скорость инициации  $\lambda$  и количество кодонов  $S$  в каждом сайте (пропорциональное размеру рибосомы). Каждый «сайт» имеет соответствующую скорость трансляции  $\lambda_i$ , которая оценивается на основании ко-адаптации между кодонами «сайта» и пулом тРНК организма. Выходные данные модели содержат вероятности размещения рибосом в каждом «сайте» и скорости трансляции, иными словами величину потока рибосом через систему. Рисунок по статье [Reuveni et al., 2011].

Модель на Рисунке 10, разработанная для изучения влияния вторичных структур на эффективность трансляции генов *S. cerevisiae*, позволила обнаружить интересные особенности данного процесса [Mao et al., 2014]. Ранее разработанные стохастические модели [Cannarozzi et al., 2010; Reuveni et al., 2011; Tuller et al., 2011; Ciandrini et al., 2013; Shah et al., 2013] также учитывают вторичные структуры в мРНК. Однако они учитывают фиксированные структуры, рассчитанные *in vitro*, и не учитывают их динамику в процессе трансляции.

Несмотря на то, что, как считается, вторичные структуры в мРНК уменьшают скорость элонгации трансляции [Wen et al., 2008; Zheng et al., 2010] в эксперименте *in vitro* была обнаружена достоверная положительная корреляция между стабильностью вторичной

структуры мРНК и количеством белка у дрожжей. Разработанная модель позволила сделать предположение, объясняющее данный эффект. В результате анализа была обнаружена значимая положительная корреляция ( $r = 0,10$ ;  $P < 4,8 \times 10^{-9}$ ) между стабильностью вторичной структуры мРНК без рибосом и плотностью рибосом на мРНК в процессе трансляции. Но, что еще более важно, была обнаружена значимая отрицательная корреляция между стабильностью вторичной структуры мРНК без рибосом и расстоянием между рибосомами в процессе трансляции ( $r = -0,81$ ;  $P < 2,20 \times 10^{-16}$ ). Это говорит о том, что более стабильная исходная вторичная структура мРНК приводит к уменьшению расстояния между рибосомами в процессе трансляции. А это в свою очередь приводит к тому, что меньше шпилек образуется на пути рибосом [Mao et al., 2014].



**Рис. 10.** Модель для симуляции процесса трансляции. (А) Первая рибосома (темно-синий цвет) встречает шпильку с одной свободной энергией. (В) третья рибосома встречает шпильку в той же позиции уже с другой свободной энергией из-за взаимодействия с ней второй рибосомы. Рисунок по статье [Mao et al., 2014].

При исследовании эффекта положения шпильки на мРНК было показано, что шпильки в 5'-районе обеспечивают большее расстояние между рибосомами в процессе элонгации, чем шпильки, расположенные ближе к 3'-концу мРНК. Тем самым уменьшается вероятность образования заторов рибосом и преждевременной остановки трансляции. Было показано, что при малых дистанциях между рибосомами в процессе элонгации трансляции фактором, определяющим эффективность данного процесса, является кодонный состав. Это

происходит из-за того, что вторичные структуры на небольших участках нуклеотидных последовательностей между рибосомами практически не образуются. С другой стороны, при больших дистанциях между рибосомами влияния вторичных структур на скорость трансляции возрастает [Mao et al., 2014].

### **Заключение к обзору литературы**

Как видно из обзора литературы, трансляция является сложным и многоступенчатым процессом. Регуляция ее эффективности осуществляется на всех уровнях.

Показано, что на скорость трансляции влияют различные контекстные характеристики нуклеотидной последовательности гена, такие как кодонный состав и вторичные структуры в мРНК. На выбор тРНК в А-сайте рибосомы могут влиять тРНК и кодон, располагающиеся в Р-сайте, посредством локальных контекстных эффектов. Не всегда выбор оптимального кодона является наилучшим. В некоторых случаях включение редких кодонов может даже повысить эффективность трансляции, благодаря уменьшению столкновений рибосом в процессе их движения вдоль мРНК. Также снижение скорости элонгации в определенных позициях на мРНК, например, кодирующих междоменные промежутки в синтезируемом полипептиде, способствует правильному фолдингу белка [Welch et al. 2009a].

Разработано несколько алгоритмов, которые позволяют исследователям манипулировать различными параметрами дизайна генов, чтобы повысить эффективность их экспрессии в целевом организме. В идеале, такие алгоритмы должны быть основаны на модели, позволяющей точно предсказывать эффективностью экспрессии по параметрам дизайна гена [Welch et al. 2009a].

Из изложенного выше вытекает необходимость в создании программного обеспечения для оценки эффективности трансляции генов на основании учета как кодонного состава, так и потенциальных вторичных структур. Программа должна позволять анализировать большое количество геномов для выявления различных особенностей их нуклеотидных последовательностей. Последующее применение данной программы для анализа открытых рамок считывания в геномах различных организмов позволит лучше изучить контекстные характеристики, влияющие на эффективность трансляции генов.

## ГЛАВА 2. МЕТОДЫ И АЛГОРИТМЫ

### 2.1. Геномные последовательности

Как было показано в разделе 1.2.1.6. главы «Обзор литературы», многоклеточным организмам присуще такое явление как тканеспецифичность. Это значит, что один и тот же ген в разных тканях организма может иметь разный уровень экспрессии и, соответственно, разную эффективность трансляции. Таким образом, тканеспецифичность может скрывать или сглаживать влияние изменений нуклеотидного состава генов на эффективность экспрессии. С другой стороны, одноклеточные организмы не обладают данной особенностью. Кроме этого одноклеточные характеризуются более высокими значениями эффективной численности популяции и скорости размножения. И поскольку все изменения, происходящие в геноме одноклеточных, напрямую отражаются на приспособленности всего организма, эволюция нуклеотидных последовательностей у них идет быстрее, чем у многоклеточных. Поэтому в качестве целевых объектов данного исследования были выбраны именно одноклеточные организмы.

В качестве исходных данных использовались геномные последовательности прокариот и одноклеточных эукариот (Таблица 3 в разделе 3.1.3. Таблицы 1-3 в Приложении) в gbk формате, скачанные из базы данных NCBI GenBank (<ftp://ftp.ncbi.nih.gov/genomes/>). В исследование брались только те организмы, у которых геном был полностью секвенирован и аннотирован. Для проведения расчетов была использована база данных GenBank по состоянию на 13 июня 2013 г.

Для каждого организма была создана отдельная директория, содержащая один или несколько gbk-файлов, соответствующих хромосомам организма. В общей сложности расчеты были проведены для 2771 генома: 2582 генома бактерий, 165 – архей и 24 – одноклеточных эукариот. Из файлов с геномами считывались только белок-кодирующие последовательности (CDS), имеющие длину кодирующей части не менее 30 кодонов и не имеющие метки «pseudo» (т.е. не являющиеся псевдогенами). Вместе с кодирующими последовательностями генов считывались их 5'- и 3'-фланкирующие районы длиной 600 нуклеотидов каждый.

### 2.2. Индекс эффективности элонгации трансляции EEI

Для оценки эффективности элонгации трансляции генов исследуемых организмов в данной работе использовался индекс эффективности элонгации трансляции EEI (Elongation Efficiency Index). EEI был разработан сотрудниками ИЦиГ СО РАН В. А. Лихошваем и Ю. Г.

Матушкиным [Лихошвай и Матушкин, 2000]. Данный индекс рассчитывается для каждого гена организма и имеет смысл средней скорости прохождения стадии элонгации трансляции.

Выбор в пользу данного индекса был сделан на основании его особенностей, перечисленных ниже.

Во-первых, для расчета  $E EI$  кроме аннотированного генома организма не требуются никакие дополнительные данные, например, выборки наиболее высокоэкспрессируемых генов. В качестве данной выборки используются гены рибосомных белков (ГРБ), которые, как известно, являются одними из высокоэкспрессирующихся в большинстве одноклеточных организмов [Sharp and Li, 1986; Владимиров, 2007].

Во-вторых, данный индекс позволяет учитывать не только кодонный состав гена, как большинство других индексов (см. раздел 1.3.1. главы «Обзор литературы»), но и его насыщенность локальными инвертированными повторами (потенциальными вторичными структурами в мРНК). Это дает возможность применять  $E EI$  при исследовании организмов, для которых показано отсутствие корреляции между кодонным составом генов и эффективностью их экспрессии или отсутствие неравномерности по использованию синонимичных кодонов.

В-третьих, данный индекс позволяет показать, какая из характеристик (кодонный состав или потенциальные вторичные структуры) оказывает основное влияние на эффективность трансляции.

Как следует из модели трансляции, представленной в разделе 1.1. главы «Обзор литературы», скорость трансляции обратно пропорциональна времени, затрачиваемому рибосомой на следующие стадии:

- 1) размещение в А-сайте изоакцепторной тРНК;
- 2) транспептидация;
- 3) транслокация.

В данном исследовании время, затрачиваемое рибосомой на стадию транспептидации, полагалось равным для всех кодонов и всех генов и поэтому не учитывалось при расчете  $E EI$ . Отсюда индекс  $E EI$  рассчитывается по следующей формуле:

$$E EI(i) = K / (w_1 T_a(i) + w_2 T_e(i)), \quad (16)$$

$i$  – номер гена,  $T_a(i)$  – параметр, имеющий смысл среднего времени размещения в А-сайте рибосомы изоакцепторной аминоксил-тРНК,  $T_e(i)$  – параметр, имеющий смысл среднего времени, затрачиваемого рибосомой на стадию транслокации,  $K$  – нормирующий множитель, обеспечивающий границы индекса от 0 до 10,  $w_1 = \{0, 1\}$  и  $w_2 = \{0, 1\}$  – индикаторные коэффициенты, определяющие учет слагаемых в значении индекса. Всего имеется три нетривиальные комбинации индикаторных коэффициентов:

- а)  $w_1 = 1, w_2 = 0$  – учитывается только слагаемое  $T_a(i)$ ;

- б)  $w_1 = 0, w_2 = 1$  – учитывается только слагаемое  $T_e(i)$ ;  
 в)  $w_1 = 1, w_2 = 1$  – учитываются оба слагаемых  $T_a(i)$  и  $T_e(i)$ .

### 2.2.1. Учет кодонного состава гена при расчете индекса ЕЕІ

Первое слагаемое  $T_a$  имеет смысл среднего времени размещения в А-сайте рибосомы изоакцепторной аминоксил-тРНК. Как было показано в разделе 1.2.1. главы «Обзор литературы», чем выше концентрация изоакцепторной аминоксил-тРНК, тем быстрее комплементарная тРНК попадает в А-сайт рибосомы. Также было показано, что концентрации тРНК пропорциональны концентрациям соответствующих кодонов в выборке высокоэкспрессирующихся генов. Таким образом, параметр  $T_a$  для каждого гена вычисляется на основании анализа его кодонного состава по следующей формуле:

$$T_a(i) = \sum_{j=1}^{n_i} \beta_{\delta(i,j)} / n_i, \quad (17)$$

$$\beta_{\delta(i,j)} = \frac{\sum_{m=1}^C \sqrt{\alpha_m}}{\sqrt{\alpha_{\delta(i,j)}}}, \quad (18)$$

где величина  $1/\beta_{\delta(i,j)}$  интерпретируется как оптимальная относительная концентрация аминоксил-тРНК, комплементарной  $j$ -ому учитываемому кодону, а  $\alpha_{\delta(i,j)}$  и  $\alpha_m$  имеют смысл частот использования кодонов  $\delta(i,j)$  и  $m$  в выделенной подвыборке генов,  $n_i$  – количество кодонов в гене  $i$ ,  $C$  – общее число кодонов (Рисунок 11). В качестве выделенной подвыборки выступает набор генов заранее заданного количества (либо численно, либо в процентах от общего числа генов в геноме организма). Изначально гены выбираются случайно, затем выборка постепенно изменяется в соответствии с рассчитываемыми значениями индекса ЕЕІ, пока не стабилизируется на конкретном оптимальном составе генов с наибольшими значениями ЕЕІ (подробнее см. раздел 2.3.3. главы «Методы и алгоритмы»).

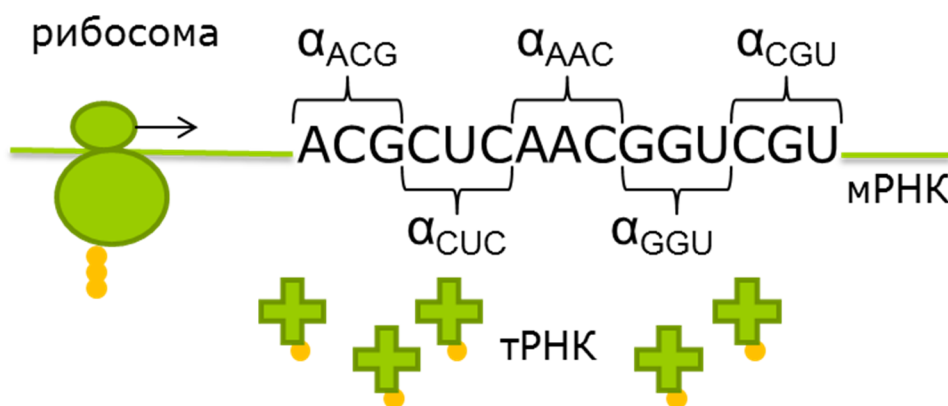


Рис. 11. Параметры модели для расчета  $\beta$  при учете кодонного состава.  $\alpha$  – частота кодона.

### 2.2.2. Учет потенциальных вторичных структур в мРНК при расчете индекса EEI

Второе слагаемое  $T_e(i)$  имеет смысл среднего времени, затрачиваемого рибосомой на стадию транслокации. Как было показано в разделе 1.2.2. главы «Обзор литературы», на скорость движения рибосомы вдоль мРНК (скорость транслокации) в процессе трансляции могут влиять шпильки, образующиеся перед рибосомой. Так как шпильки образуются за счет спаривания комплементарных нуклеотидов, поиск в мРНК таких комплементарных участков (локальных инвертированных повторов) позволяет оценить вероятность образования вторичных структур. Таким образом, параметр  $T_e(i)$  вычисляется на основе оценки самокомплементарности  $i$ -ой мРНК по следующей формуле:

$$T_e(i) = t_{\min}(1 - p(i)) + t_{\max}p(i), \quad (19)$$

где  $t_{\min}$  – минимальное условное время транслокации,  $t_{\max}$  – максимальное условное время транслокации,  $p(i)$  – вероятность реализации максимального условного времени транслокации. Чем выше самокомплементарность мРНК, тем выше значение вероятности  $p(i)$ , и наоборот. Формула для расчета  $p(i)$  следующая (Рисунок 12):

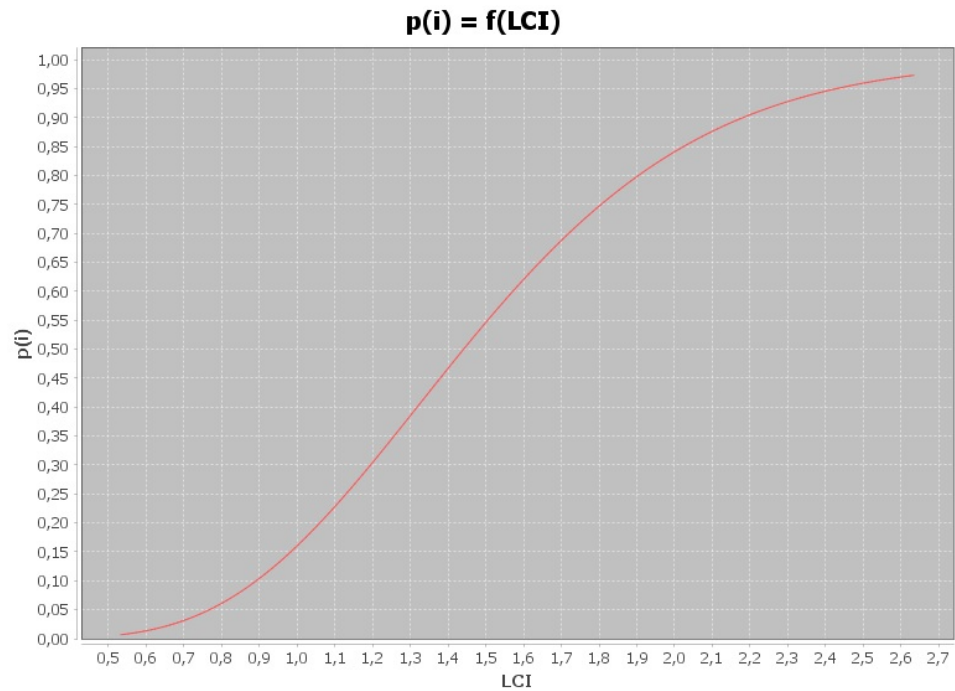
$$p(i) = \int_0^{LCI(i)} \frac{k^{n+1}x^n}{G(n+1)} e^{-kx} dx, \quad (20)$$

$$k = m/\sigma^2, \quad (21)$$

$$n = (m/\sigma)^2, \quad (22)$$

где  $m$  и  $\sigma^2$ , соответственно, математическое ожидание и дисперсия положительной случайной величины, имеющей плотность распределения  $\frac{k^{n+1}x^n}{G(n+1)} e^{-kx}$ , где  $G(n+1)$  – Гамма-функция,  $LCI(i)$  – индекс локальной комплементарности, учитывающий самокомплементарность мРНК (см. раздел 2.2.2.1. главы «Методы и алгоритмы»). Следует отметить, что результаты существенно не изменяются, если в качестве  $p(i)$  выбирать другие формы S-образной зависимости от аргумента  $LCI(i)$  [Likhoshvai and Matushkin 2002].

Параметр  $LCI$  в зависимости от его вида оценивает либо среднее количество локальных инвертированных повторов в мРНК, либо среднюю энергетическую стабильность шпилек, образуемых данными повторами (см. раздел 2.2.2.1. главы «Методы и алгоритмы»). Чем больше повторов в мРНК, либо чем стабильнее потенциальные вторичные структуры, тем выше значение параметра  $LCI$ . Из Формулы 20 видно, что чем выше  $LCI$ , тем больше  $p(i)$  – вероятность того, что рибосома замедлит свое движение вдоль мРНК. Так учитывается вклад локальной комплементарности мРНК в скорость элонгации трансляции.



**Рис. 12.** График зависимости  $p(i)$  от  $LCI(i)$ .

### 2.2.2.1. Индекс локальной комплементарности LCI

Индекс локальной комплементарности отражает насыщенность нуклеотидной последовательности мРНК локальными инвертированными повторами. Данные повторы потенциально могут образовать шпильки в мРНК.

Время, затрачиваемое рибосомой на расплетание шпилек может быть пропорционально как их количеству, так и их стабильности. Поэтому при учете влияния вторичных структур на время элонгации трансляции рассматриваются оба этих варианта. Это отражено в использовании двух типов индекса LCI: без учета энергии шпилек (LCI1) и с учетом энергии (LCI2). Первый тип основан на предположении, что рибосома последовательно расплетает вторичную структуру независимо от ее свободной энергии. Второй тип предполагает, что время, затрачиваемое рибосомой на расплетание вторичной структуры, зависит от свободной энергии этой структуры.

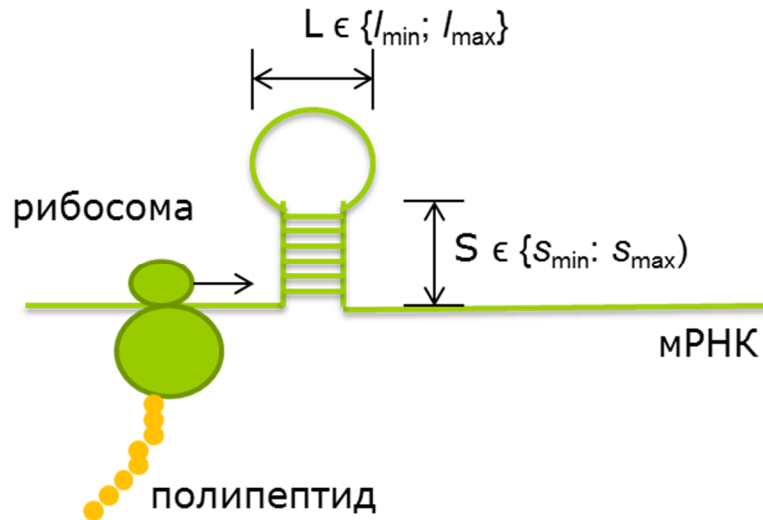
Кроме этого учитываются только локальные вторичные структуры размером не больше  $\approx 60$  нуклеотидов (изменяемый параметр). Рибосомы при движении по мРНК блокируют формирование шпилек. Это приводит к тому, что вторичные структуры могут образовываться только на участках мРНК между двумя движущимися друг за другом на стадии элонгации трансляции рибосомами (как это описано в разделе 1.3.3. главы «Обзор литературы»). Размер данного участка в процессе трансляции может значительно изменяться от нуля до нескольких сотен нуклеотидов [Mao et al., 2014]. В среднем можно принять размер такого окна равным примерно 100 нуклеотидам.



Индекс LCI1 (без энергии) рассчитывается по следующей формуле:

$$LCI1(i) = \frac{\sum_{m=1}^{m_i - s_{\max} - l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[ \sum_{l=l_{\min}}^{l_{\max}} \zeta(\text{con}(m, m+s-1), \overline{\text{con}(m+s+l, m+2s+l-2)}) \right] \right\}}{m_i - s_{\max} - l_{\max}}, \quad (23)$$

где  $\text{con}(i, j)$  – слово с  $i$ -го по  $j$ -й нуклеотид и  $\overline{\text{con}(i, j)}$  – комплементарное слово с  $j$ -го по  $i$ -й нуклеотид ( $i \leq j$ ).  $\zeta(\text{con1}, \text{con2}) = 1$ , если слова  $\text{con1}$  и  $\text{con2}$  идентичны, иначе  $\zeta(\text{con1}, \text{con2}) = 0$ ,  $i = [1, N_{\text{CDS}}]$ , где  $N_{\text{CDS}}$  – общее число считанных из файла с геномом CDS. Длина совершенного повтора (размер стебля шпильки) не меньше  $s_{\min} = 3$  и не больше  $s_{\max} = 6$ , расстояние между повторами (длина петли шпильки) не меньше  $l_{\min} = 3$  и не больше  $l_{\max} = 50$  (Рисунок 13) (значения данных параметров получены эмпирически и соответствуют наилучшему соотношению качество/производительность). Значение  $LCI1(i)$  имеет смысл среднего числа комплементарных нуклеотидов, приходящихся на один нуклеотид анализируемой последовательности.



**Рис. 13.** Параметры модели при расчете LCI для учета потенциальных вторичных структур в мРНК.  $S$  – длина локального инвертированного повтора (длина стебля потенциальной шпильки),  $L$  – расстояние между локальными инвертированными повторами (размер петли потенциальной шпильки).

Индекс LCI2 (с энергией) рассчитывается по формуле:

$$LCI2(i) = \frac{\sum_{m=1}^{m_i - s_{\max} - l_{\max}} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[ \sum_{l=l_{\min}}^{l_{\max}} \psi(\text{con}(m, m+s), \overline{\text{con}(m+s+l-1, m+2s+l-2)}) \right] \right\}}{m_i - s_{\max} - l_{\max}}, \quad (24)$$

где  $\psi$  – энергия вторичной структуры [Turner and Sugimoto, 1988]. Остальные обозначения те же, что и для LCI1.  $s_{\min} = 3$ ,  $s_{\max} = 6$ ,  $l_{\min} = 3$ ,  $l_{\max} = 50$ . Значение  $LCI2(i)$  имеет смысл средней стабильности вторичных структур в мРНК.

Энергия вторичной структуры  $\psi$  рассчитывается следующим образом. Сначала рассчитывается энергия стебля шпильки. Она состоит из суммы энергий взаимодействия формирующих ее динуклеотидов. Значения энергий определены экспериментально и берутся из статьи [Turner and Sugimoto, 1988]. После этого на основании длины петли  $L$  из статьи

[Turner and Sugimoto, 1988] берется значение ее энергии. К первому значению прибавляется второе, и, если результат меньше нуля, такая шпилька считается стабильной и учитывается в расчетах. Если результат больше или равен нулю, такая петля считается нестабильной и не учитывается в расчетах.

### 2.2.2.2. Индекс локальной комплементарности индивидуального нуклеотида $LCI(i, j)$

Индекс  $LCI(i, j)$  используется при построении профилей средних  $LCI$  по всем генам организма, для выявления распределения стабильности вторичных структур в разных районах мРНК. Как было показано в разделе 1.2.2. главы «Обзор литературы», основное влияние на скорость элонгации оказывают шпильки в 5'-районе мРНК. Подробное исследование данного района, а также 3'-района мРНК при помощи индекса локальной комплементарности индивидуальных нуклеотидов позволяет обнаружить у некоторых организмов характерные особенности (см. разделы 3.2.2.-3.2.3. главы «Результаты и обсуждение»).

$LCI(i, j)$  рассчитывается для каждого нуклеотида отдельно в отличие от обычных индексов  $LCI1$  и  $LCI2$ , которые рассчитываются для всей последовательности мРНК целиком. Данный индекс показывает среднюю энергию потенциальных вторичных структур, в образовании которых принимает участие конкретный нуклеотид. Формула для расчета индекса  $LCI(i, j)$  следующая:

$$LCI(i, j) = \sum_{m: m \leq j \leq m+s-1} 10 \left\{ \sum_{s=s_{min}}^{s_{max}} \left[ \sum_{l=l_{min}}^{l_{max}} -\psi(\text{con}(m, m+s-1), \overline{\text{con}(m+s+l, m+2s+l-2)}) \right] \right\}, \quad (25)$$

где  $i$  – номер CDS,  $j$  – порядковый номер нуклеотида. Остальные обозначения такие же, как для  $LCI2$ .

Профили со средними значениями индекса  $LCI$  индивидуальных нуклеотидов по всем генам *S. cerevisiae* хорошо согласуются с экспериментальными данными, полученными методом PARS [Kertesz et al., 2010] (см. раздел 1.2.2.3. главы «Обзор литературы»). Экспериментальный и теоретический профили очень похожи. Также очень близки значения коэффициентов корреляции и достоверностей. Коэффициент корреляции между плотностью рибосом вдоль транскрипта (экспериментальная эффективность элонгации трансляции мРНК) и значением PARS score (экспериментальная вероятность шпильки в районе старт-кодона) равен  $r = -0,1$  ( $P < 10^{-4}$ ) в позиции -10 (точка с наибольшей достоверностью коэффициента корреляции). Корреляция между EEI (теоретическая эффективность трансляции) и  $LCI(i, j)$  (теоретическая стабильность шпильки) равна примерно -0,055 ( $P < 6,64 \times 10^{-5}$ ) в позиции -19 (точка с наибольшим значением коэффициента корреляции). Такое высокое согласие с экспериментальными данными говорит о хорошей предсказательной способности индекса  $LCI(i, j)$ .

### 2.2.3. Пять типов EEI и определение типа, лучше всего оценивающего эффективность элонгации трансляции в исследуемом организме

Как было отмечено в разделе 2.2. главы «Методы и алгоритмы», всего имеется три нетривиальные комбинации индикаторных коэффициентов  $w_1$  и  $w_2$  в формуле расчета EEI. Также в расчетах используется два типа индекса LCI (LCI1 и LCI2). В итоге из Формулы 16 получаются пять типов индекса EEI:

- 1) EEI1 =  $K/T_a$  – учитывается только кодонный состав гена;
- 2) EEI2 =  $K/T_e(LCI1)$  – учитывается только количество вторичных структур в мРНК;
- 3) EEI3 =  $K/T_e(LCI2)$  – учитывается только стабильность вторичных структур в мРНК;
- 4) EEI4 =  $K/(T_a+T_e(LCI1))$  – учитываются и кодонный состав, и количество вторичных структур в мРНК;
- 5) EEI5 =  $K/(T_a+T_e(LCI2))$  – учитываются и кодонный состав, и стабильность вторичных структур в мРНК.

Таким образом, для каждого гена исследуемого организма рассчитывается пять типов индекса EEI1-EEI5. Для определения, какой из типов индекса лучше всего оценивает эффективность элонгации трансляции в конкретном организме, в списках генов, отсортированных по каждому из пяти типов индекса EEI, выделяются гены рибосомных белков (ГРБ) и рассчитываются их ранг (M1-M5) и стандартное отклонение (R1-R5) по формулам:

$$M = \frac{1}{N_{rib}} \sum_{i=1}^{N_{rib}} x_i, \quad (26)$$

$$R = \sqrt{\frac{1}{N_{rib}} \sum_{i=1}^{N_{rib}} (M - x_i)^2}, \quad (27)$$

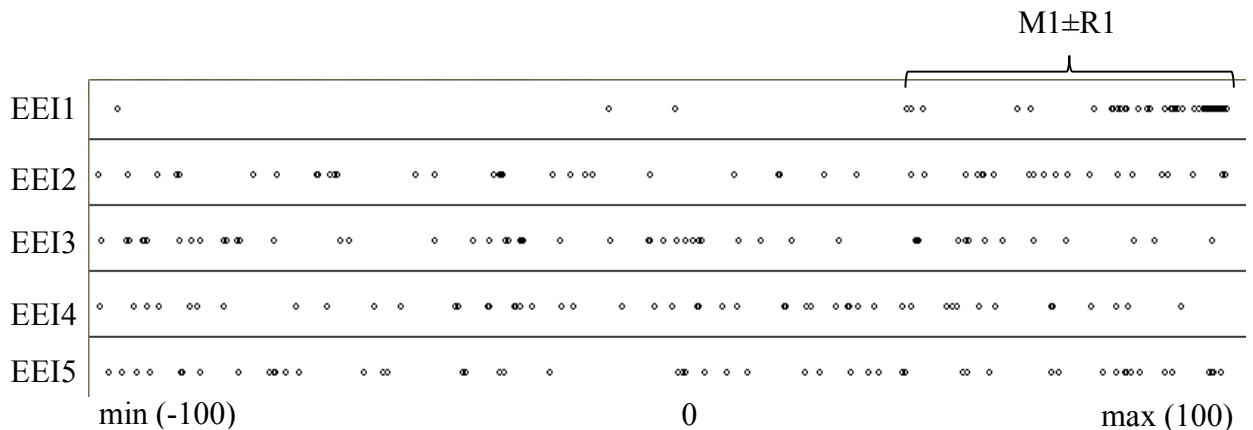
где  $N_{rib}$  – количество ГРБ,  $x_i$  – позиция ГРБ в отсортированном по увеличению индекса EEI списке генов. Для удобства параметры M и R нормируются таким образом, чтобы их значения лежали в интервалах [-100; 100] и [0; 100] соответственно.

ГРБ выбраны в качестве маркеров высокоэкспрессирующихся генов, так как для большинства одноклеточных организмов известно, что гены рибосомных белков характеризуются высоким уровнем экспрессии. Стоит отметить, что при выборе в качестве маркеров других высокоэкспрессирующихся генов, в том числе экспериментально определенных, получаются аналогичные результаты [Vladimirov et al., 2007].

Для каждого организма считается основным тот тип индекса EEI, для которого параметр M принимает наибольшее значение (из пяти M1-M5), а параметр R – наименьшее (из пяти R1-R5), т.е. ГРБ больше смещены в сторону высоких значений EEI и расположены плотнее друг к другу (Рисунок 14). Таким образом, если эти гены получают высокие значения EEI, то это можно считать критерием адекватности данного индекса в оценке эффективности их

экспрессии. Достоверности ранга и стандартного отклонения оценивались для организмов с количеством генов <1000 методом Монте-Карло, а для остальных методами теории вероятностей на основании Центральной предельной теоремы [Владимиров и др., 2007]. Практически для всех исследованных организмов уровень достоверности  $P < 10^{-3}$ . Низкие значения достоверности характерны для организмов, для которых значения ранга  $M < 30$ . Эти организмы названы в данной работе особенными и описаны в разделе 3.1.4. главы «Результаты и обсуждение».

Например, у *Herpetosiphon aurantiacus* DSM 785 основным является 1-ый тип индекса EEI, так как он показывает более высокий уровень экспрессии ГРБ, чем остальные четыре типа индекса (Рисунок 14). Принадлежность организма к тому или иному типу EEI говорит о типе эволюционной оптимизации первичной структуры его генов (см. раздел 2.3.1. главы «Методы и алгоритмы»).



**Рис. 14.** Схема расположения рибосомных генов (черные кружки) среди других генов (не отмечены) организма (*Herpetosiphon aurantiacus* DSM 785), упорядоченных по увеличению индекса EEI. ГРБ расположены правее и плотнее в первом типе EEI.

Далее под параметрами  $M$  и  $R$  для конкретного организма будут подразумеваться наибольшее значение из ( $M1-M5$ ) и соответствующее ему значение из ( $R1-R5$ ). При наличии двух и более равных наибольших значений  $M$  выбирается то, которому соответствует наименьшее значение  $R$ .

## 2.3. Программа EloE

### 2.3.1. Общие сведения о программе EloE

Для проведения исследований в рамках данной работы было создано специальное программное обеспечение EloE (Elongation Efficiency), реализующее описанный выше алгоритм расчета индекса EEI. Данная программа в виде веб-приложения доступна по ссылке <http://www-bionet.sccc.ru:7780/EloE/>. Также на программу EloE было получено авторское свидетельство

№2014662021, зарегистрированное 19.11.2014, «Программа для автоматической оценки эффективности элонгации трансляции генов различных организмов (EloE)», авторы: Соколов В.С., Зураев Б.С., Генаев М.А.

Первоначально разработанный В. А. Лихошваем и Ю. Г. Матушкиным алгоритм был реализован на языке программирования Fortran. Программа представляла собой комплекс отдельных приложений, каждое из которых требовалось отдельно настраивать и последовательно запускать. Затем Н. В. Владимиров объединил при помощи Perl-скрипта данный комплекс программ в одну. Это дало возможность запускать всего одну программу один раз для расчета одного организма. Также Н. В. Владимировым программа была размещена в открытом доступе по адресу <http://www.mgs2.bionet.nsc.ru/mgs/programs/eei-calculator/> (на данный момент эта версия программы убрана из общего доступа).

Несмотря на модернизацию, у программы оставался еще ряд недостатков, таких как:

а) невозможность обсчета нескольких организмов за один запуск. Данное требование сейчас особенно актуально из-за большого объема имеющихся данных по секвенированным геномам различных организмов;

б) невысокая скорость работы за счет того, что программа разбита на отдельные последовательно запускаемые модули;

в) сложность в изменении стартовых параметров расчета.

Поэтому в рамках данной работы был заново реализован алгоритм расчета индекса EEI, опробованного и хорошо себя зарекомендовавшего. Реализация была произведена на языке программирования Java. Все вышеперечисленные недостатки удалось либо уменьшить, либо полностью ликвидировать:

а) программа EloE позволяет обсчитывать геномы сразу нескольких организмов за один запуск;

б) производительность возросла в несколько раз;

в) стало проще изменять входные параметры расчета.

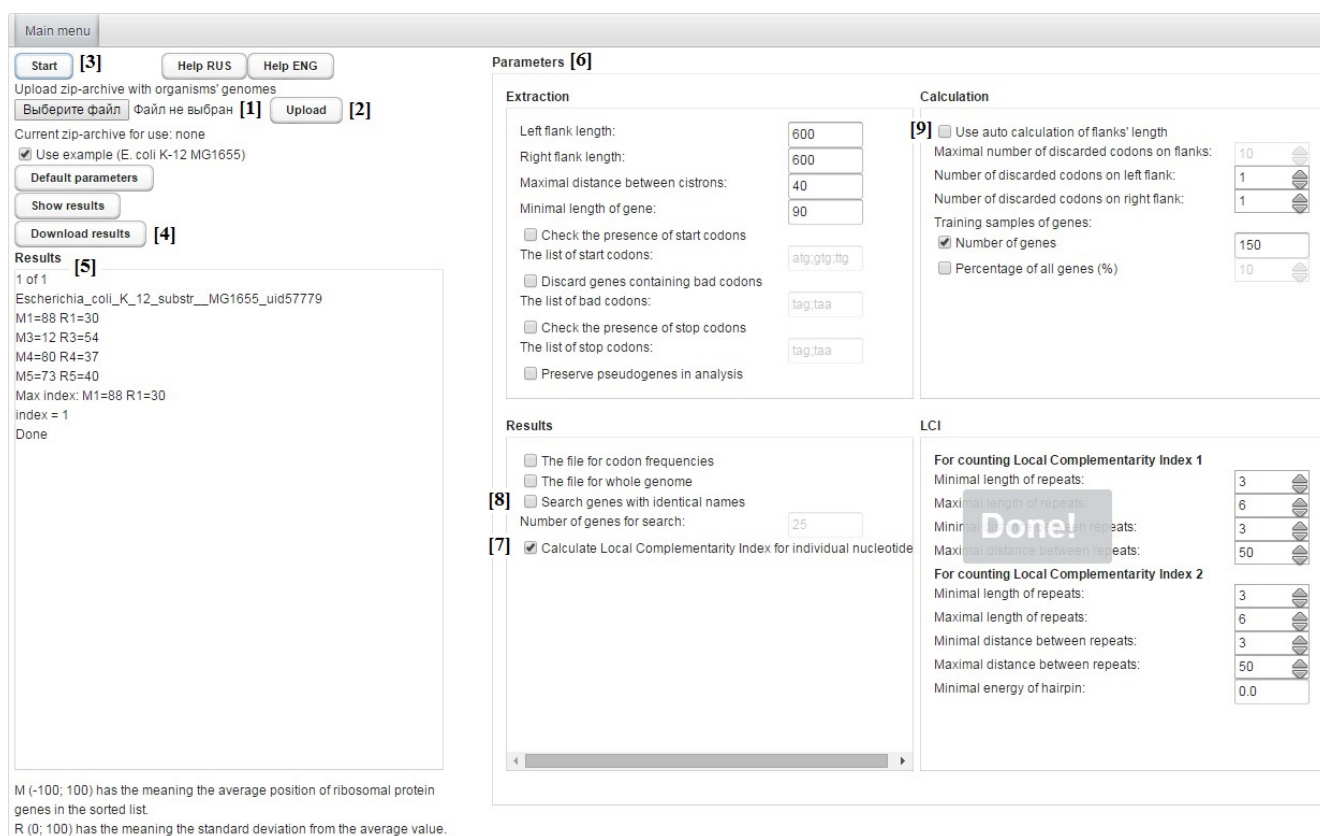
Кроме этого было добавлено несколько новых функций, в том числе построение графиков с LCI профилями индивидуальных нуклеотидов.

На Рисунке 15 показан интерфейс программы EloE. Программа позволяет определять, в каком направлении прошла массовая оптимизация первичной структуры генов организма в процессе эволюции:

1) индекс EEI1 – в организме в процессе эволюции прошла массовая оптимизация первичной структуры генов в сторону отбора предпочтительных/быстрых кодонов, для уменьшения временных и энергетических затрат на кодон-антикодонное взаимодействие мРНК с изоакцепторной тРНК;

2) индексы EEI2/EEI3 – отбор шел на минимизацию количества/энергии вторичных структур в мРНК, для уменьшения энергетических затрат на денатурацию шпилек перед рибосомным комплексом в процессе движения последнего вдоль мРНК на стадии элонгации трансляции;

3) индексы EEI4/EEI5 – в организме отбор шел как в сторону оптимизации кодового состава гена, так и в сторону оптимизации вторичных структур в мРНК.



**Рис. 15.** Графический интерфейс веб-приложения EloE (скриншот интерфейса программы EloE). Кнопки Выберите файл [1], Upload [2], Start [3] и Download results [4] в верхнем левом углу позволяют запускать приложение и загружать файлы с результатами. В поле Results [5] видна часть результатов обсчета генома *E. coli* K-12 MG1655. Как можно видеть, в данном организме работает первый тип индекса EEI1. Область Parameters [6] позволяет задавать параметры расчета, в том числе [7]-[9] (подробнее см. текст).

Второй основной функцией разработанной программы является сортировка нуклеотидных последовательностей генов исследуемого организма в порядке уменьшения их вычисленной эффективности прохождения стадии элонгации трансляции. Так как высокоэкспрессируемые гены характеризуются высокой эффективностью как стадии транскрипции, так и трансляции, данный список в первом приближении дает информацию о высокоэкспрессируемых генах исследуемого организма.

Далее по тексту под высоко(низко)экспрессируемыми генами подразумеваются гены организма с наибольшими (наименьшими) значениями индекса эффективности элонгации трансляции EEI, т.е. наибольшей (наименьшей) теоретической эффективностью трансляции.

Кроме описанных функций программа EloE позволяет:

- а) проводить массовый анализ геномов сразу нескольких организмов;
- б) проводить гибкий анализ генома организма за счет большого количества изменяемых параметров расчета (Рисунок 15, [6]);
- в) для прокариотических организмов определять положение гена в опероне, если он не является в нем единственным;
- г) строить профили средних значений LCI индексов индивидуальных нуклеотидов для всех генов организма и для отдельных специальных групп (например, только для первых цистронов в оперонах) (Рисунок 15, [7]);
- д) при одновременном анализе 2-10 организмов искать гены с одинаковыми идентификаторами («именами») и отображать их расположение на графиках для каждой пары организмов (Рисунок 15, [8]); это позволяет быстро обнаружить наличие у пары организмов генов с одинаковыми «именами», но разной эффективностью элонгации трансляции (и, возможно, экспрессии). Это упрощает анализ организмов, для которых нет данных о семействах ортологичных генов.

На расчет индексов EEI на примере генома *E.coli* программа EloE тратит порядка 20 секунд. В это время не входит расчет дополнительных параметров, таких как LCI индивидуальных нуклеотидов. Соответственно, на геномы меньших размеров уходит меньше времени. Например, на обработку генома *Mycoplasma fermentans* затрачивается примерно 5 секунд. А на большие геномы может требоваться от одной минуты до нескольких десятков минут. На *S. cerevisiae* требуется около 30 секунд. При расчете LCI индивидуальных нуклеотидов затрачиваемое время увеличивается примерно в два-три раза.

### 2.3.2. Входные и выходные данные программы EloE

Исходными данными являются файлы с аннотированной нуклеотидной последовательностью полного генома организма в формате gbk (файлы могут быть взяты в базе данных GenBank <ftp://ftp.ncbi.nih.gov/genbank/genomes>). Для произведения расчетов требуется создание zip архива с аннотированными геномами (gbk) исследуемых организмов. Геном каждого организма должен располагаться в архиве в отдельной папке. Загрузка архива в программу осуществляется с помощью кнопки «Upload» (Рисунок 15, [2]). Все результаты, в том числе список генов организма, отсортированных по индексу EEI, сохраняются в файлы и могут быть загружены после окончания расчетов (кнопка «Download results» (Рисунок 15, [4])).

Основные файлы с результатами располагаются для каждого организма в отдельной папке «Organism\_name»:

- 1) «organism\_name\_all.txt» – файл со всеми пятью типами индекса EEI, рассчитанными для всех генов организма, учитываемых в расчетах;
- 2) «organism\_name\_eeiN.txt» ( $N = \{1, 2, 3, 4, 5\}$ ) – файл только с тем типом индекса EEI, который работает в данном организме;
- 3) «organism\_name\_genes\_and\_flanks.txt» – файл с подробной информацией по каждому гену и его нуклеотидной последовательностью с флангами;
- 4) «organism\_name\_number\_eei.txt» – файл с указанием номера гена, его позиции в опероне (только для прокариот) и значения EEI;
- 5) «organism\_name\_gibpos.txt» – файл с расположением генов рибосомных белков в списках генов организма, отсортированных по каждому из пяти типов EEI в порядке увеличения (Рисунки 16-17).

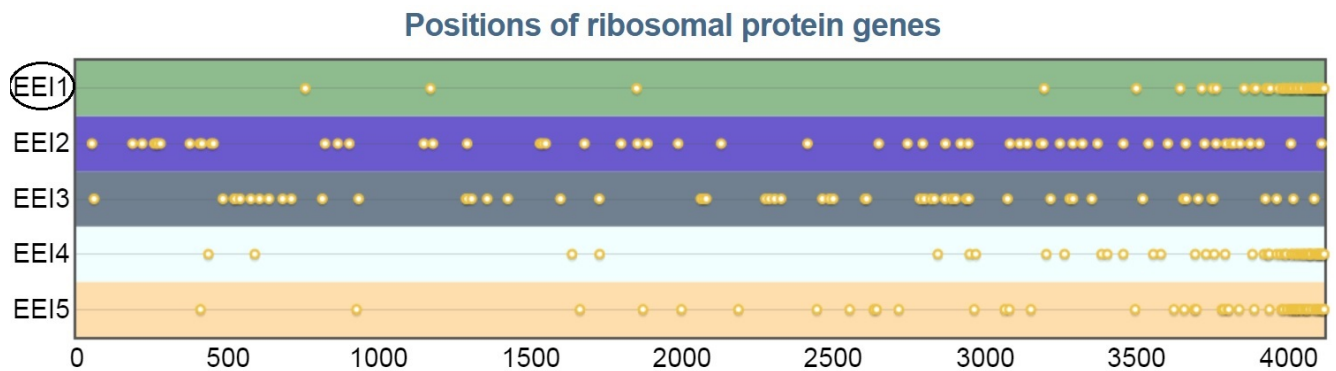
Общие результаты по всем анализируемым геномам собраны в одном файле «organism\_index.txt». Данные во всех файлах разделены знаком табуляции.

Интерфейс программы позволяет изменять параметры расчетов: размеры фланкирующих районов генов, длины инвертированных повторов в мРНК и расстояние между мономерами этих повторов. В начале и на конце гена обычно располагаются специфические кодоны, характерные именно для сайтов начала/конца трансляции. Поэтому их учет может негативно повлиять на расчеты EEI. В программе можно указать количество кодонов на краях гена, которые не будут учитываться в расчетах, или поставить галочку «Use auto calculation of flanks' length» (Рисунок 15, [9]). Тогда программа сама определит оптимальное количество не учитываемых кодонов. Также можно заказать дополнительные выходные файлы.

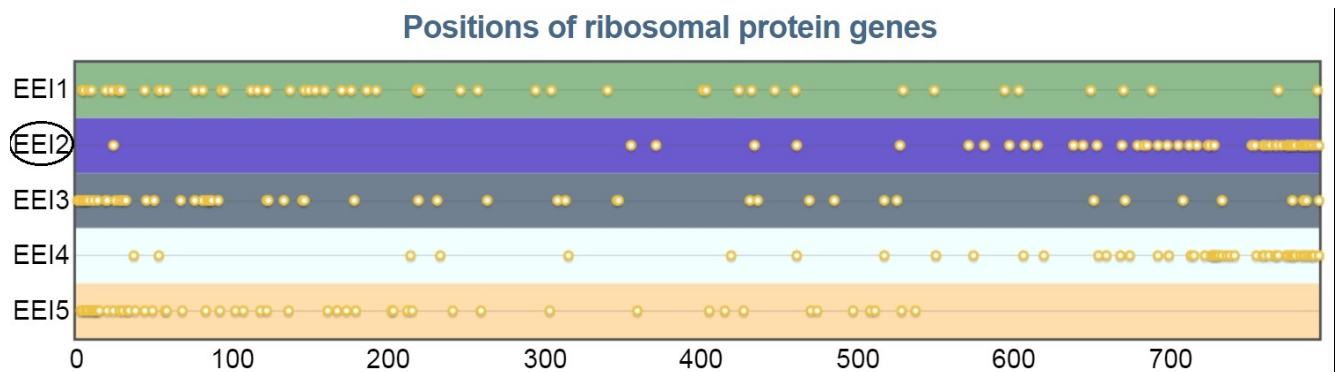
Одной из возможностей программы является генерация файла «organism\_name\_lciij\_profile\_out.xlsx» (при установленной галочке «Calculate Local Complementarity Index for individual nucleotides» (Рисунок 15, [7])). В нем хранятся значения для построения профилей индексов локальной комплементарности индивидуальных нуклеотидов, которые строятся в веб-приложении. LCI индивидуальных нуклеотидов имеет смысл средней стабильности шпилек, которые потенциально могут быть образованы локальными совершенными инвертированными повторами определенной длины в мРНК. Данные шпильки могут замедлять движение рибосомного комплекса в процессе элонгации трансляции [Lopinski et al., 2000; Takyar et al., 2005]. Расчет LCI индивидуальных нуклеотидов производится в районах старт-кодона ( $\pm 600$  нуклеотидов) и стоп-кодона ( $\pm 600$  нуклеотидов) трансляции. Для этого вместе с последовательностью гена из файла gbk экстрагируются фланкирующие районы длиной 600 нуклеотидов. Изменение длины экстрагируемых флангов не влияет на расчеты индексов EEI.



Для каждого организма EIoE строит график с позициями ГРБ для каждого типа EEI (Рисунки 16-17). Гены на графике отсортированы в порядке увеличения EEI. Таким образом, наилучший тип EEI для организма – это тот тип, для которого гены рибосомных белков располагаются правее и плотнее. Как видно из Рисунка 16 в *E. coli* лучше всего работает первый тип EEI, т.е. эффективность элонгации трансляции в большей степени зависит от частот кодонов в гене. У *Mycoplasma fermentans* JER (Рисунок 17) лучше работает второй тип EEI, т.е. эффективность элонгации трансляции в основном определяется количеством локальных инвертированных повторов в гене.

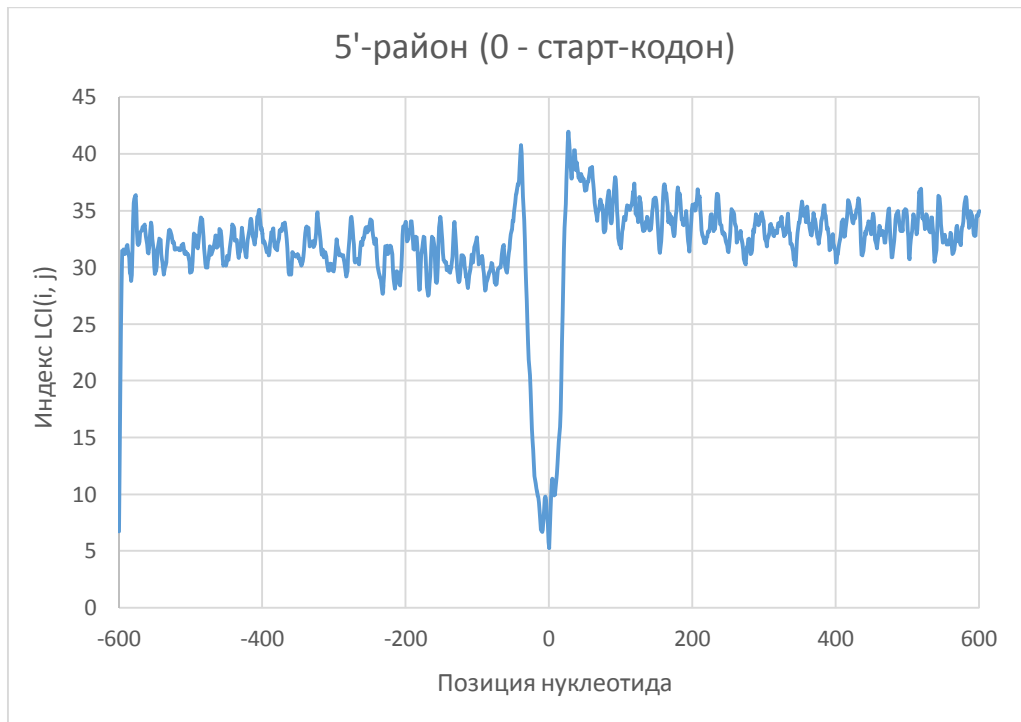


**Рис. 16.** Распределение генов рибосомных белков (точки) в списке генов *E. coli* K-12 MG1655, расположенных слева-направо в порядке увеличения EEI1-5 (скриншот интерфейса программы EIoE). Ось абсцисс – позиция гена в списке. Наилучший тип индекса (EEI1) выделен кружком – гены рибосомных белков расположены правее и плотнее.

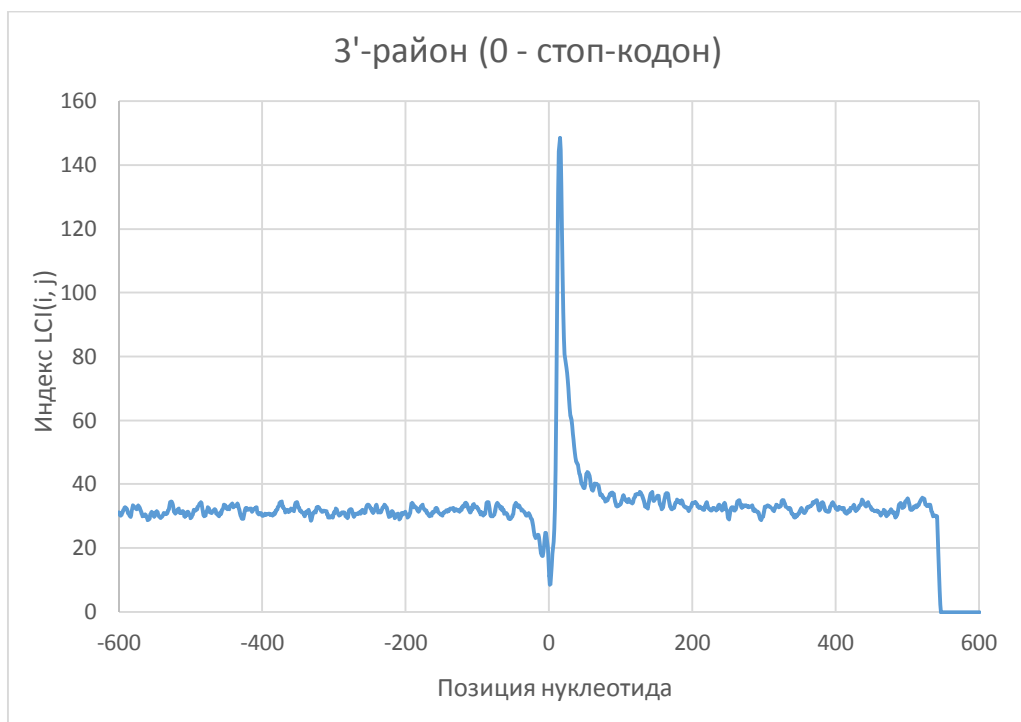


**Рис. 17.** Распределение генов рибосомных белков (точки) в списке генов *Mycoplasma fermentans* JER, расположенных слева-направо в порядке увеличения EEI1-5 (скриншот интерфейса программы EIoE). Ось абсцисс – позиция гена в списке. Наилучший тип индекса (EEI2) выделен кружком – гены рибосомных белков расположены правее и плотнее.

Приведенные на Рисунках 18-19 профили LCI индивидуальных нуклеотидов отображают среднюю по всем генам организма стабильность потенциальных шпилек в районах старт- и стоп-кодонов трансляции. Спад профиля в районе старт-кодона (Рисунок 19) говорит о меньшей стабильности шпилек в данном районе мРНК. С другой стороны, пик профиля в районе стоп-кодона (Рисунок 19) говорит о более высокой стабильности шпилек.



**Рис. 18.** Визуализация профиля среднего значения LCI индивидуальных нуклеотидов по всем генам *E. coli* K-12 MG1655. Ноль на оси абсцисс – старт-кодон.



**Рис. 19.** Визуализация профиля среднего значения LCI индивидуальных нуклеотидов по всем генам *E. coli* K-12 MG1655. Ноль на оси абсцисс – стоп-кодон трансляции.

Выходные файлы содержат такие параметры генов как: EEI, LCI, GC-состав, длина, позиция в опероне (для прокариот) и др. Дополнительной функцией программы является построение усредненных профилей стабильности вторичных структур в районах 5'- и 3'-концов мРНК всех генов организма. Для прокариот можно выбирать по каким генам строить профиль:

по всем или только по первым, внутренним, последним или единственным цистронам в оперонах.

### 2.3.3. Алгоритм работы программы EIoE

Алгоритм работы программы EIoE представлен на Рисунке 20.



**Рис. 20.** Алгоритм работы программы EIoE. Во время итерационного цикла программа автоматически определяет выборку генов, которые характеризуются наибольшей эффективностью элонгации трансляции.

После считывания нуклеотидных последовательностей из файлов с геномами организмов, для каждой последовательности рассчитываются частоты кодонов и количество или стабильность потенциальных вторичных структур. Затем случайным образом выбирается обучающая выборка из  $N = 150$  последовательностей. Такой размер выборки выбран из

соображений о примерно одинаковом количестве генов «домашнего хозяйства» в одноклеточных организмах. Данный параметр можно изменять, указывая как конкретное число генов в выборке, так и процент от полного количества генов в геноме. После этого идет итерационный алгоритм расчета пяти типов индекса EEI. На каждой итерации заново выбирается обучающая выборка, состоящая из последовательностей с наибольшими значениями соответствующего индекса EEI. Алгоритм прекращает свою работу после того, как обучающая выборка перестает меняться. Обычно это происходит за 5-10 итераций. В результате получаются пять списков с генами организма, каждый из которых отсортирован по одному из пяти типов индекса EEI.

#### **2.4. Статистический анализ**

При получении описанных ниже результатов статистическая обработка данных и их визуализация были проведены с помощью языка программирования R и программ RStudio и Microsoft Excel. Для расчета корреляций в работе использовались коэффициенты корреляции Пирсона и Спирмена. Также при анализе для оценки различий между исследуемыми выборками использовались t-критерий Стьюдента, критерий Манна-Уитни,  $\phi$ -критерий Фишера. На нормальность выборки проверялись при помощи теста Шапиро-Вилка.

## ГЛАВА 3. РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

### 3.1. Исследование геномов одноклеточных организмов при помощи программы EюЕ. Расчет индексов эффективности элонгации трансляции

#### 3.1.1. Исследование геномов бактерий

При помощи программы EюЕ был проведен анализ 2582 геномов бактерий, представленных в базе данных (БД) NCBI GenBank по состоянию на 13 июня 2013 года. Новые данные, внесенные в БД GenBank, существенно не влияют на полученные результаты. Поэтому приведенные графики можно считать актуальными. На Рисунке 21 представлен график распределения исследованных бактерий по пяти типам индекса EEI.

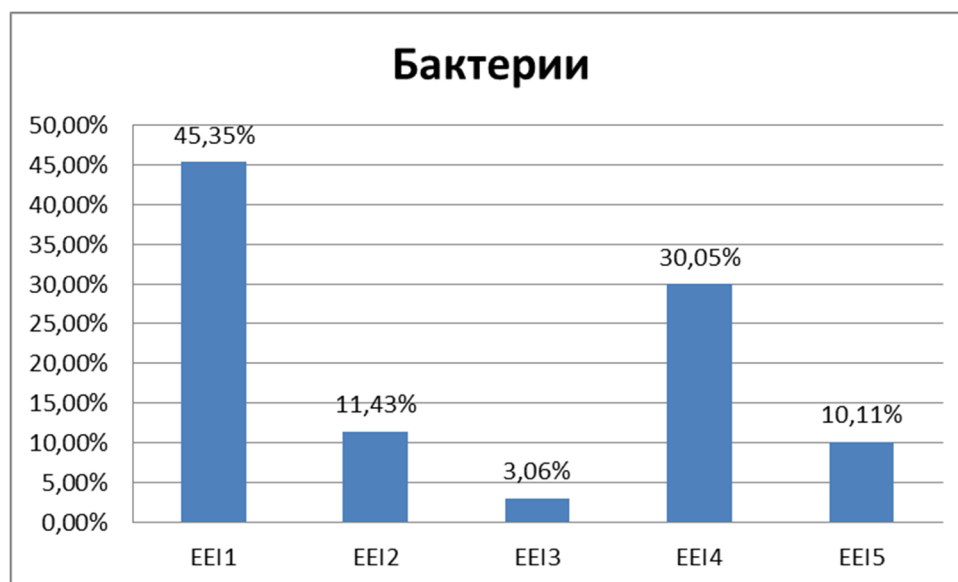


Рис. 21. Распределение 2582 организмов бактерий по пяти типам индекса EEI.

В Таблице 1 приведено несколько представителей каждого типа индекса EEI. Все результаты представлены в Таблице 1 в Приложении.

Таблица 1. Некоторые представители пяти типов индекса EEI среди бактерий (синим цветом выделены наибольшие значения параметра M).

Организм	Тип EEI	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
<i>Corynebacterium pseudotuberculosis</i> FRC41	1	99	0	-43	55	-41	57	84	30	50	58
<i>Vibrio cholerae</i> M66-2		97	3	5	67	7	60	81	31	68	50
<i>Escherichia coli</i> str. K-12 substr. MDS42		92	16	15	64	15	54	84	28	79	33
<i>Arcobacter</i> sp. I	2	-31	60	81	21	-25	57	77	36	-33	55
<i>Mycoplasma hyorhinis</i> MCLD		-24	58	79	22	-23	60	75	27	-29	56
<i>Spiroplasma taiwanense</i> CT-1		-51	56	75	41	-36	59	55	53	-47	59
<i>Azoarcus</i> sp. BH72	3	-19	57	67	28	70	29	58	39	64	33
<i>Desulfobulbus propionicus</i> DSM2032		-65	38	47	46	66	42	-24	70	25	49
<i>Nitrosomonas europaea</i> ATCC19718		-53	60	8	71	49	51	-20	70	15	54
<i>Arthrobacter arilaitensis</i> REI17	4	86	13	21	54	28	52	92	9	87	26
<i>Sulfurimonas autotrophica</i> DSM16294		13	53	43	48	-16	61	89	18	-14	57
<i>Legionella pneumophila</i> subsp. Pneumophila str. Philadelphia I		18	56	29	58	-12	57	70	39	-8	55
<i>Serratia plymuthica</i> 4RX13	5	-27	36	54	41	72	25	87	17	93	9
<i>Pseudomonas putida</i> H8234		76	29	51	37	68	25	74	27	90	19
<i>Lactobacillus delbrueckii</i> subsp. Bulgaricus ATCC11842		66	50	-10	66	50	50	9	60	75	48

Как видно из графика на Рисунке 21, большинство бактерий относится к первому типу индекса EEI – 45,35%. У данных организмов эффективность элонгации трансляции в большей степени зависит от частот кодонов в генах, чем от вторичных структур в мРНК. Типичный представитель – *Escherichia coli* K-12.

Вторая по размеру группа организмов (30,05%) относится к 4-му типу EEI. В данном случае эффективность элонгации трансляции определяется как кодонным составом, так и количеством вторичных структур в мРНК. Типичный представитель – *Sulfurimonas autotrophica* DSM16294.

На третьем месте по численности организмы, относящиеся ко 2-му типу EEI. Эффективность элонгации трансляции их генов определяется только количеством вторичных структур в мРНК. Типичный представитель – *Mycoplasma hyorhinis* MCLD.

В 10,11% бактерий лучше всего работает 5-ый тип EEI. У данных организмов важную роль в эффективности элонгации трансляции играют кодонный состав гена и энергетическая стабильность потенциальных вторичных структур в мРНК. Типичный представитель – *Pseudomonas putida* H8234.

Самая малочисленная группа бактерий (3,06%) относится к 3-му типу EEI. Эффективность элонгации трансляции определяется энергетической стабильностью потенциальных вторичных структур в мРНК. Типичный представитель – *Azoarcus* BH72.

### 3.1.2. Исследование геномов архей

Всего было проанализировано 165 геномов архей, имеющих в БД NCBI GenBank по состоянию на 13 июня 2013 года. График распределения организмов по типам индекса EEI представлен на Рисунке 22. В Таблице 2 приведено несколько представителей каждого типа индекса EEI. Все результаты представлены Таблице 2 в Приложении.

Из графика на Рисунке 22 видно, что среди архей, в отличие от бактерий, нет организмов, принадлежащих к 3-ему и 5-ому типам индекса EEI. По-видимому, археи лишены энергозависимого механизма расплетания вторичных структур в мРНК.

Большинство организмов (58,79%) принадлежит к 4-му типу индекса EEI. Это говорит о существенной роли кодонного состава и количества вторичных структур в мРНК в определении эффективности элонгации трансляции. Типичный представитель – *Methanococcus voltae* A3.

Вторая по размеру группа организмов принадлежит 2-му типу EEI – 23,64%. Эффективность элонгации у данных организмов определяется в основном количеством вторичных структур в мРНК. Типичный представитель данной группы – *Methanothermus fervidus* DSM 2088.

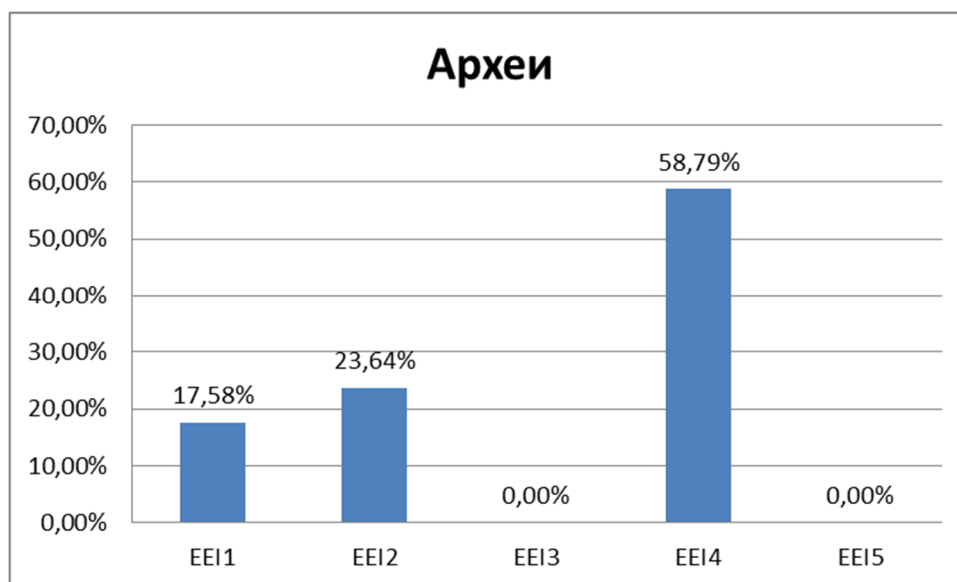


Рис. 22. Распределение 165 организмов архей по пяти типам индекса EEI.

Таблица 2. Некоторые представители трех типов индекса EEI среди архей (синим цветом выделены наибольшие значения параметра M).

Организм	Тип EEI	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
<i>Thermococcus kodakarensis</i> KOD1	1	88	21	-10	67	-55	43	63	45	9	69
<i>Halogeometricum borinquense</i> DSM11551		73	44	-3	51	-26	57	60	49	11	66
<i>Haloarcula hispanica</i> ATCC33960		62	48	-10	55	-24	53	53	46	12	63
<i>Picrophilus torridus</i> DSM9790	2	-41	57	78	26	-36	63	34	63	-50	53
<i>Methanothermus fervidus</i> DSM2088		-3	67	74	30	1	62	46	59	-8	61
<i>Caldisphaera lagunensis</i> DSM15908		-22	58	69	35	-24	61	37	63	-32	57
<i>Methanobrevibacter ruminantium</i> M1	4	10	51	42	48	-15	61	89	7	-10	56
<i>Methanococcus voltae</i> A3		-32	45	69	33	-37	51	89	19	-43	45
<i>Methanocaldococcus fervens</i> AG86		9	58	75	34	-27	65	89	23	-22	63

Последняя группа, содержащая 17,58% исследованных архей, принадлежит к 1-му типу индекса EEI. У данной группы организмов эффективность элонгации трансляции зависит только от кодонного состава генов. Типичный представитель – *Thermococcus kodakarensis* KOD1.

### 3.1.3. Исследование геномов одноклеточных эукариот

В результате работы было проанализировано 24 генома одноклеточных эукариот. График распределения организмов по типам индекса EEI представлен на Рисунке 23. Большинство организмов (62,50%) относятся к первому типу EEI. Типичный представитель – *Saccharomyces cerevisiae* S288C. *Encephalitozoon cuniculi* GB-M1 – организм с наименьшим значением параметра M1 = 22 среди данных организмов.

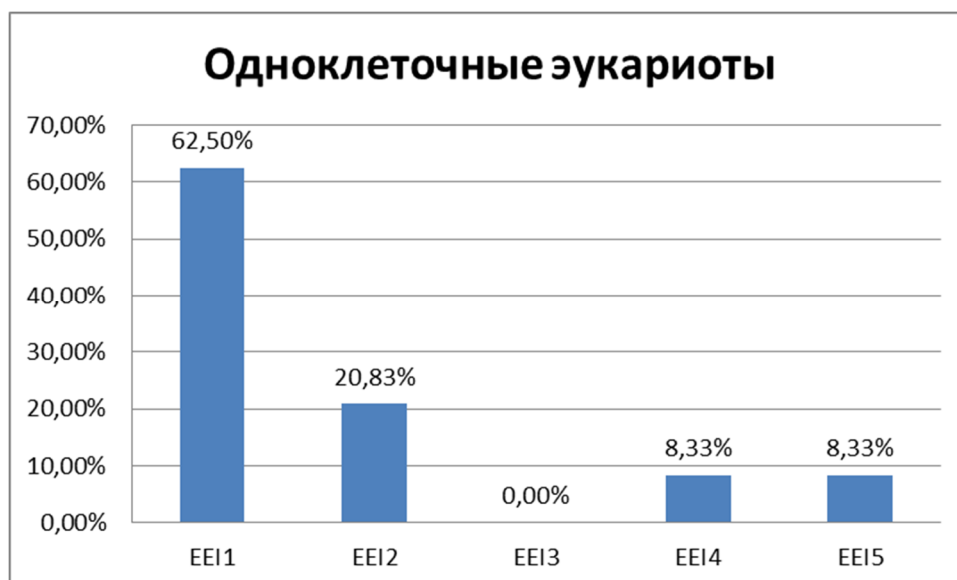


Рис. 23. Распределение 24 одноклеточных эукариот по пяти типам индекса EEI.

В Таблице 3 представлены результаты по всем 24 одноклеточным эукариотам.

Таблица 3. Средние смещения генов рибосомных белков в сторону высокоэкспрессирующихся для 24 одноклеточных эукариот (синим цветом выделены наибольшие значения параметра М, красным – организм с наибольшим параметром М < 30).

Организм	Тип EEI	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
<i>Lachancea thermotolerans</i> cbs 6340	1	99	0	-44	55	-14	52	97	3	97	20
<i>Zygosaccharomyces rouxii</i> cbs 732	1	98	1	-28	60	-51	52	52	51	34	65
<i>Tetrapisispora phaffii</i> cbs 4417	1	97	18	35	47	-37	56	95	18	-45	41
<i>Debaryomyces hansenii</i> cbs767	1	97	20	8	53	-38	60	96	20	0	69
<i>Kluyveromyces lactis</i> nrrl y-1140	1	97	21	-50	57	-52	57	88	28	46	67
<i>Torulaspora delbrueckii</i>	1	97	21	-67	44	-34	60	88	25	70	55
<i>Komagataella pastoris</i> gs115	1	96	15	-2	59	-40	63	90	19	-8	73
<i>Naumovozyma dairenensis</i> cbs 421	1	95	25	20	50	-59	52	94	25	-60	44
<i>Schizosaccharomyces pombe</i> 972h-	1	94	21	43	53	-60	48	91	25	-68	37
<i>Scheffersomyces stipitis</i> cbs 6054	1	93	23	-35	63	-32	61	47	55	20	70
<i>Yarrowia lipolytica</i> clib122	1	92	30	-16	68	-14	58	78	36	65	56
<i>Saccharomyces cerevisiae</i> S288C	1	91	34	-23	59	-35	55	86	35	57	60
<i>Cryptococcus gattii</i> wm276	1	81	34	-40	67	-51	54	75	32	30	59
<i>Cryptococcus neoformans</i> var neoformans jec21	1	78	32	-30	71	-45	60	75	30	32	64
<i>Encephalitozoon cuniculi</i> gb-m1	1	22	51	14	62	-14	61	19	51	-9	58
<i>Paramecium tetraurelia</i> strain D4-2	2	-38	44	94	5	8	59	46	47	-26	42
<i>Cryptomonas paramecium</i>	2	-50	33	62	20	3	62	-29	55	-37	42
<i>Plasmodium falciparum</i> 3d7	2	-62	34	54	46	-19	65	19	44	-42	46
<i>Guillardia theta</i>	2	-28	34	41	26	10	55	32	51	-20	37
<i>Encephalitozoon intestinalis</i> ATCC50506	2	-38	52	38	51	-23	63	-3	54	-43	54
<i>Candida dubliniensis</i> cd36	4	11	44	8	54	-42	53	94	23	-33	53
<i>Cyanidioschyzon merolae</i> strain 10d	4	31	54	18	61	6	67	51	49	29	63
<i>Ashbya gossypii</i> ATCC10895	5	-16	30	-25	55	20	48	-17	49	94	30
<i>Ostreococcus lucimarinus</i> cce9901	5	3	48	17	49	21	52	28	49	35	51

### 3.1.4. Особенности организмов, выявленные в результате анализа геномов одноклеточных организмов программой EIoE

В процессе компьютерного исследования геномов было выявлено 73 организма, у которых параметр М (наибольшее из пяти значений M1-M5), характеризующий смещение ГРБ



в сторону высокоэкспрессирующихся генов, было менее 30 ( $M < 30$ ). В данных особых организмах вероятность получить такие ранги по случайным причинам достаточно высока ( $P > 10^{-3}$ ). Поэтому ни один из типов индекса EEI не позволяет достоверно определить эффективность элонгации трансляции генов в данных организмах. В Таблице 4 представлены некоторые особые организмы с параметром  $M < 30$  (полные данные приведены в Таблице 3 в Приложении) и вероятности получить такие значения  $M$  по случайным причинам.

**Таблица 4.** Некоторые особые организмы с параметром  $M < 30$  (синим цветом выделены наибольшие значения параметра  $M$ , красным – организмы с наибольшим параметром  $M < 30$ ).

Организм	Тип EEI	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5	Вероятность
<i>Wolbachia</i> endosymbiont of <i>Culex quinquefasciatus</i> Pel	1	21	55	10	59	11	58	20	49	19	55	8,2E-03
<i>Jonesia denitrificans</i> DSM20603	1	15	32	-38	53	-15	62	-16	52	0	57	5,9E-02
<i>Oscillatoria acuminata</i> PCC6304	2	9	56	14	56	-19	55	12	57	-28	52	5,4E-02
<i>Prochlorococcus marinus</i> str. MIT9313	2	-12	50	5	61	-26	58	-16	50	-24	56	5,1E-01
<i>Sulfobacillus acidophilus</i> DSM110332	3	8	65	10	60	16	55	-23	57	15	64	3,8E-02
<i>Mycoplasma haemofelis</i> str. Langford 1	3	-17	57	-16	60	3	69	-28	52	-1	69	7,1E-01
<i>Synechococcus</i> sp. JA-3-3AB	4	16	51	13	48	-4	58	20	46	1	61	1,2E-02
<i>Candidatus carsonella ruddii</i> CS isolate Thao2000	4	0	48	10	39	7	59	13	57	-1	48	1,3E-01
<i>Candidatus hodgkinia cicadicola</i> DSEM	5	0	49	18	37	19	46	14	47	20	50	2,6E-02
<i>Mycobacterium tuberculosis</i> Casnit204	5	-25	54	14	54	4	58	10	59	18	65	2,0E-02

## 3.2. Подробное исследование организмов, принадлежащих к роду *Mycoplasma*

### 3.2.1. Распределение исследованных штаммов *Mycoplasma* по пяти типам индекса EEI

Одними из особых организмов были несколько видов рода *Mycoplasma*: *M. haemocanis*, *M. haemofelis*, *M. suis* и др. (см. Таблицу 4 в Приложении). Одновременно с этим среди микоплазм есть много видов, для которых индекс EEI хорошо работает, т.е. ГРБ достоверно определяются как высокоэкспрессируемые. Было проведено более подробное исследование организмов, принадлежащих к роду *Mycoplasma*, для выяснения причин появления особых видов.

Дополнительным основанием к проведению анализа организмов, принадлежащих именно к роду *Mycoplasma*, послужил размер их геномов. Поскольку большинство микоплазм являются паразитами, их геномы значительно редуцированы, что позволяет изучить их более детально и, возможно, более явно показать присущие им закономерности. Также изучение особенностей трансляции у *Mycoplasma* может иметь практическое значение в медицине, т.к. некоторые виды микоплазм являются паразитами человека и домашних животных (Таблица 5). А в свете того, что на данный момент процесс трансляции и механизмы его регуляции у микоплазм мало изучены, в том числе экспериментально, данное исследование является весьма актуальным.

Таблица 5. Организмы, на которых паразитируют различные виды *Mycoplasma*.

Хозяин	Паразит	Поражаемые клетки
Человек	<i>M. genitalium</i>	Реснитчатый эпителий дыхательных и половых путей
	<i>M. pneumoniae</i>	Реснитчатый эпителий трахей
Кошки	<i>C. M. haemominutum</i>	Зрелые эритроциты
	<i>M. haemofelis</i>	
	<i>C. M. turicensis</i>	
Собаки	<i>C. M. haematoparvum</i>	
	<i>M. haemocanis</i>	
Овцы и козы	<i>M. ovis</i>	
Свиньи	<i>M. suis</i>	
Коровы	<i>M. wenyonii</i>	
Ламы и альпаки	<i>C. M. haemolamae</i>	

При помощи программы EloE были проанализированы геномы 62 штаммов (под штаммом подразумевается чистая культура бактерий, изолированная от одного источника в определенное время.), принадлежащих к 27 видам *Mycoplasma*. Часть результатов представлена в Таблице 6 (все результаты представлены в Таблице 4 в Приложении).

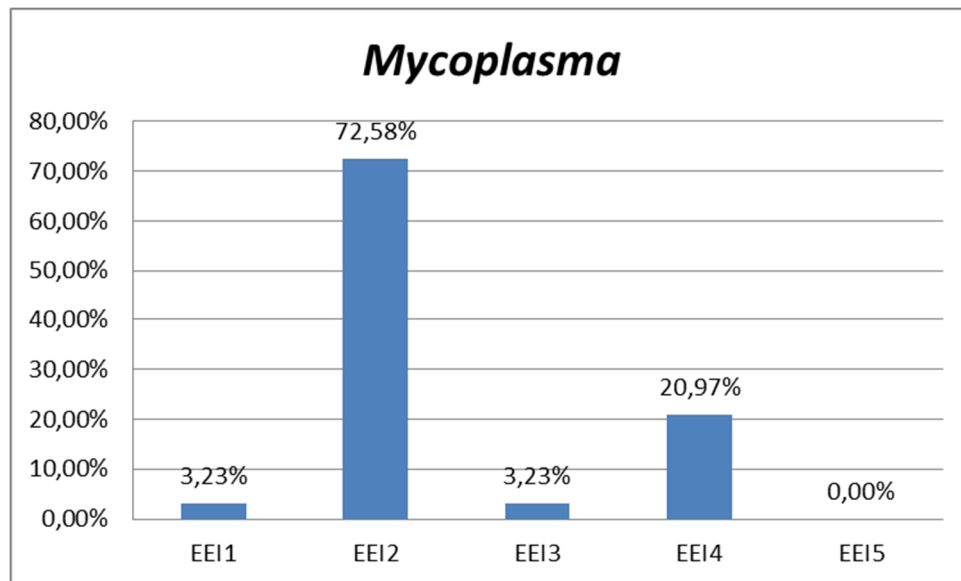
**Таблица 6.** Типы индексов для штаммов *Mycoplasma*. Синим цветом в строке выделен тот тип индекса EEI, который работает в соответствующем штамме (наибольшее значение параметра M в строке). Красным выделены штаммы с наименьшим смещением ГРБ в сторону высокоэкспрессирующихся генов ( $M \leq 30$ ). Для штаммов с одинаковыми значениями параметра M для разных типов индекса в скобках дополнительно приведены значения параметра R.

Штамм	Тип EEI	M1 (R1)	M2 (R2)	M3 (R3)	M4 (R4)	M5 (R5)
<i>Mycoplasma wenyonii</i> str. Massachusetts	1	70	-9	-21	69	25
<i>Candidatus Mycoplasma haemolamae</i> str. Purdue	1	30	5	20	25	29
<i>Mycoplasma capricolum</i> subsp. capricolum ATCC 27343	2	-60	78	-26	40	-52
<i>Mycoplasma fermentans</i> JER	2	-40	77	-42	71	-62
<i>Mycoplasma mycoides</i> subsp. capri	2	-53	76	-27	25	-51

<b>LC str. 95010</b>						
<i>Mycoplasma leachii</i> PG50	2	-56	75	-31	41	-56
<i>Mycoplasma cynos</i> C142	2	-39	74	-29	47	-47
<i>Mycoplasma synoviae</i> 53	2	-37	71	-8	50	-32
<i>Mycoplasma penetrans</i> HF 2	2	-1	71	-32	60	-37
<i>Mycoplasma hominis</i> ATCC 23114	2	-7	69	-28	67	-38
<i>Mycoplasma putrefaciens</i> KS1	2	-20	68	-21	59	-41
<i>Mycoplasma pulmonis</i> UAB CTIP	2	-13	68	5	48	-16
<i>Mycoplasma hyopneumoniae</i> 168	2	-67	68	-28	-33	-60
<i>Mycoplasma genitalium</i> G37	2	-66	59	-39	-40	-64
<i>Mycoplasma gallisepticum</i> CA06 2006 052-5-2p	2	23	49	-7	47	-1
<i>Mycoplasma suis</i> KI3806	2	18	27	0	23	3
<i>Mycoplasma suis</i> str. Illinois	2	6	26	0	17	2
<i>Mycoplasma pneumoniae</i> M129-B7	2	-10	25	-27	23	-30
<i>Mycoplasma pneumoniae</i> FH	2	-13	24	-32	21	-31
<i>Candidatus Mycoplasma haemominutum</i> Birmingham 1	2	-8	24	2	8	-9
<i>Mycoplasma haemocanis</i> str. Illinois	2	-15	-6 (63)	-6 (70)	-19	-24
<i>Mycoplasma haemofelis</i> Ohio2	3	-23	-21	7	-36	-3
<i>Mycoplasma haemofelis</i> str. Langford 1	3	-25	-16	3	-35	-12
<i>Mycoplasma agalactiae</i>	4	26	64	-24	66	-26
<i>Mycoplasma bovis</i> PG45	4	13	63	-40	66	-35
<i>Mycoplasma pneumoniae</i> 309	4	-9	23	-26	25	-29
<i>Mycoplasma pneumoniae</i> M129	4	-10	22	-25	25	-29

Суммарное распределение штаммов по типам индекса представлено на Рисунке 24. Как можно видеть, в большинстве штаммов работает второй тип индекса EEI2 – эффективность элонгации трансляции зависит только от количества вторичных структур в мРНК и не зависит от кодонного состава генов. Почти у всех микоплазм ГРБ хорошо определяются как высокоэкспрессируемые, значения ранга М для них высоки. Но есть виды со значительно более

низкими значениями параметра  $M$ : *C. M. haemominutum*, *M. suis*, *M. pneumoniae* и, особенно, *M. haemocanis* и *M. haemofelis* (Таблица 6, красный цвет). У данных видов практически отсутствует смещение ГРБ в сторону высокоэкспрессирующихся.



**Рис. 24.** Распределение 62 штаммов микоплазм по типам индекса EEI.

Объяснить этот факт можно с помощью двух альтернативных предположений:

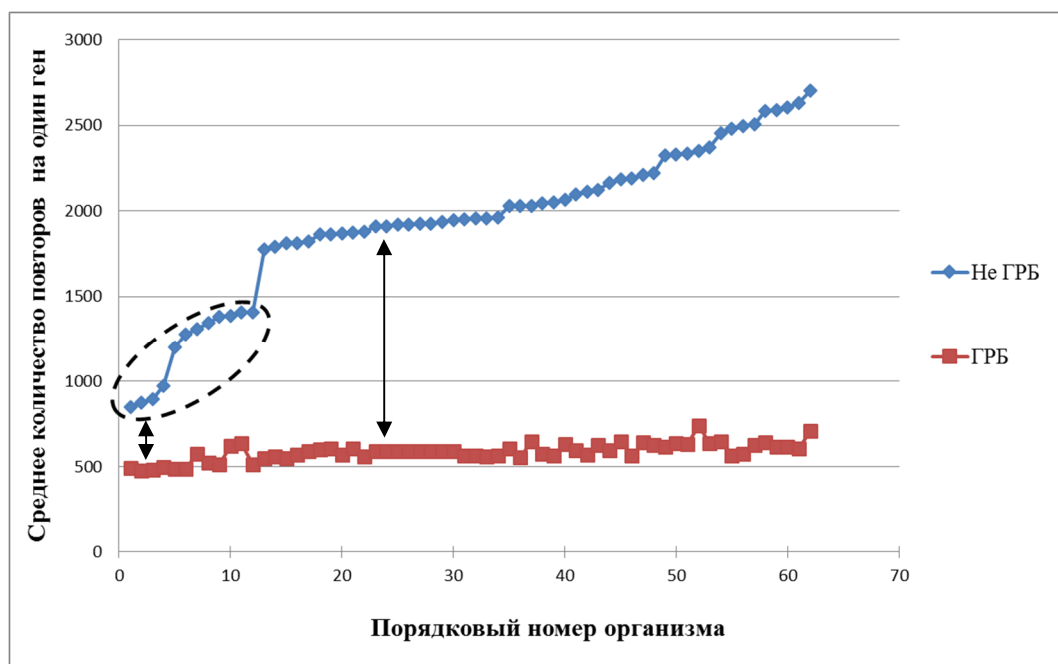
1. ГРБ у данных видов не являются высокоэкспрессируемыми;
2. все остальные гены у данных видов характеризуются повышенным уровнем экспрессии, что нивелирует уровень экспрессии ГРБ.

### 3.2.2. Анализ количества совершенных локальных инвертированных повторов в генах различных штаммов *Mycoplasma*

Для выбора одного из двух объяснений наличия микоплазм с низкими значениями  $M$ , были рассчитаны средние значения количества совершенных локальных инвертированных повторов на один ген для всех 62 штаммов *Mycoplasma*. Полученные результаты были отсортированы по увеличению среднего количества локальных инвертированных повторов на один нерибосомный ген (ген, не являющийся ГРБ – не ГРБ) и отображены на графике, представленном на Рисунке 25. Аналогичные результаты получаются, если количество повторов дополнительно нормировать на длину каждого гена.

Из Рисунка 25 видно, что ГРБ у всех штаммов мало отличаются друг от друга по количеству локальных инвертированных повторов, что хорошо согласуется с предположением о высокой консервативности нуклеотидных последовательностей данных генов. С другой стороны, не ГРБ у разных штаммов могут значительно различаться по среднему количеству повторов на один ген. В частности, интересующие нас «особые» штаммы с низкими

значениями параметра  $M$  (выделены красным цветом в Таблице 6) попали в самую крайнюю группу с наименьшими значениями количества повторов на один ген (ограничена пунктиром на Рисунке 25). Благодаря этому разница между ГРБ и остальными генами у данных штаммов значительно меньше, чем у других (размер стрелочек на Рисунке 25), что и объясняет их особенность.



**Рис. 25.** Среднее число совершенных локальных инвертированных повторов на один ген для каждого штамма. Пунктиром выделены штаммы с параметром  $M \leq 30$ . Стрелочки показывают разницу между средним количеством повторов на один рибосомный и нерибосомный ген.

Установленный факт неравномерности по среднему числу локальных инвертированных повторов в генах разных штаммов *Mycoplasma* является новым и ранее неизученным, поэтому требует более подробного исследования.

У четырех штаммов (*M. haemofelis* Ohio2, *M. haemofelis* Langford1, *M. haemocanis* Illinois, *C. M. haemolamae* Purdue) с самыми низкими значениями среднего количества локальных инвертированных повторов на один нерибосомный ген (крайние слева на Рисунке 25) были рассмотрены 100 генов с самым низким количеством локальных инвертированных повторов для определения их функций. Большинство из данных 100 генов отмечены как «hypothetical protein» и их функции не известны. Но среди генов с известными функциями встречаются следующие: белки рибосом, субъединицы ДНК полимераз, субъединицы АТФ-синтаз, ферменты из комплекса аминоксил-тРНК-синтаз, трансмембранные переносчики. Таким образом, самое низкое содержание локальных инвертированных повторов наблюдается в генах «домашнего хозяйства».

### 3.2.3. Филогенетический анализ исследуемых *Mycoplasma*

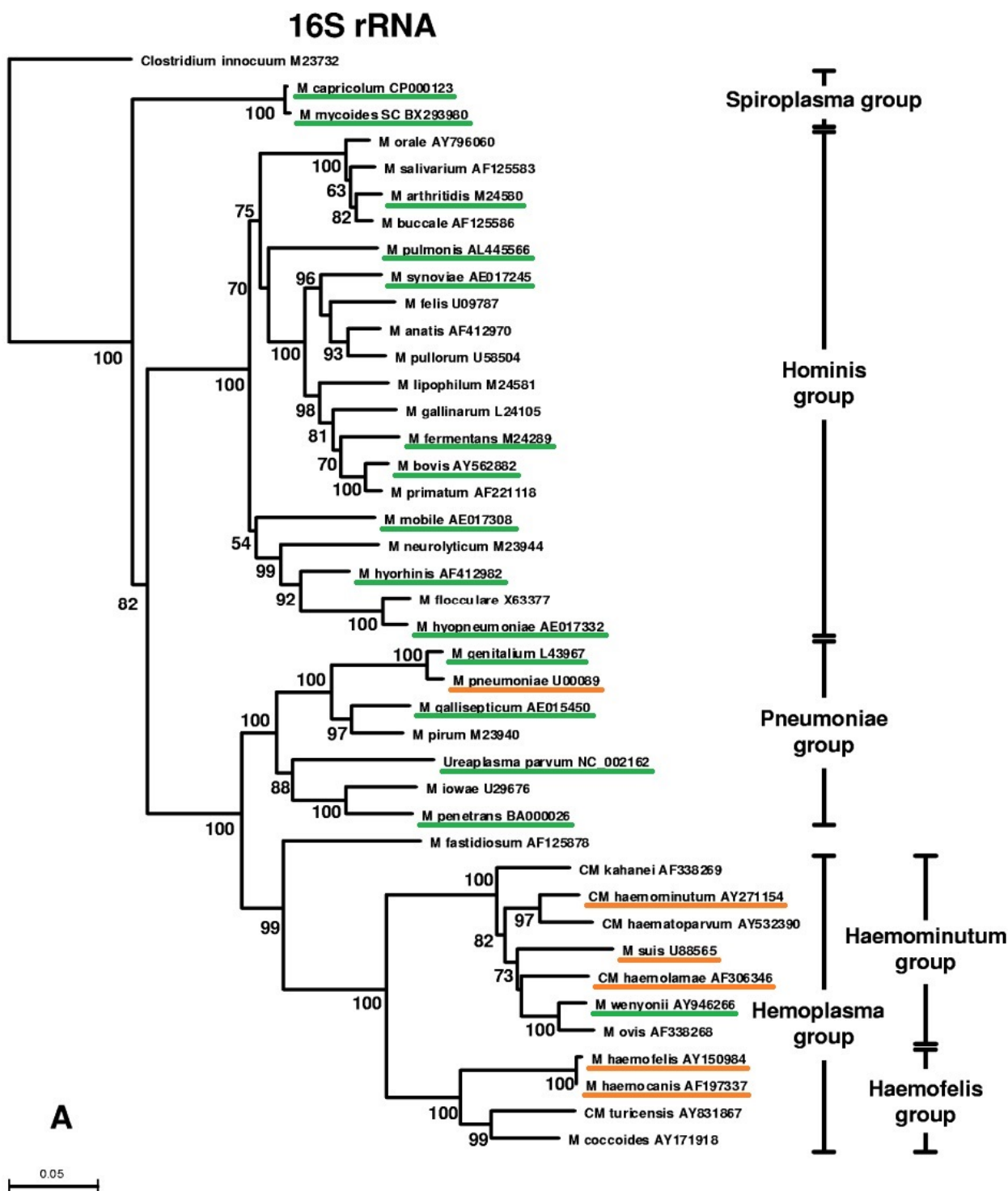
Чтобы понять, почему же именно у данных штаммов (ограничены пунктиром на Рисунке 25) наблюдается такое низкое значение среднего количества локальных инвертированных повторов на один ген, был проведен филогенетический анализ исследуемых *Mycoplasma*.

Филогенетическое дерево *Mycoplasma*, построенное на основе анализа последовательностей 16S рРНК, было взято из статьи [Peters et al., 2008] и представлено на Рисунке 26. На дереве зеленым цветом отмечены виды со значением параметра  $M > 30$ , а оранжевым – с  $M \leq 30$ . Видно, что почти все виды, отмеченные оранжевым, (кроме *M. pneumoniae*) попадают в группу гемоплазм. Достоверность преобладания особых видов микоплазм с параметром  $M < 30$  в группе гемоплазм была рассчитана при помощи  $\phi$ -критерия Фишера. Было выделено две группы организмов: «гемоплазмы» и «не гемоплазмы». После этого проверялась гипотеза о различии частот встречаемости видов с параметром  $M < 30$  в каждой из групп. Показано преобладание особых микоплазм в группе гемоплазм с достоверностью более 99% ( $P < 0,01$ ).

Гемоплазмы – это гемотрофные организмы, их жизнь связана с красными кровяными тельцами (эритроцитами). Считается, что эти микоплазмы паразитируют на поверхности эритроцитов и даже могут проникать внутрь них.

На основании всего вышеизложенного можно предположить, что в особых условиях обитания на поверхности или внутри эритроцитов данные виды эволюционировали в сторону уменьшения количества совершенных локальных инвертированных повторов (потенциальных вторичных структур) в генах. Возможно, таким образом они уменьшили энергетические затраты на процесс трансляции, чтобы повысить эффективность их экспрессии.

Вместе с другими микоплазмами в группу гемоплазм попала *M. wenyonii*, паразит крупного рогатого скота. У данного организма тоже отмечается низкое значение среднего количества повторов на один ген (пунктирная область на Рисунке 25). Однако в нем работает первый тип индекса EEI и значение параметра  $M = 70$ , т.е. ГРБ располагаются в области высокоэкспрессирующихся генов. Это говорит о том, что данный организм эволюционировал в сторону оптимизации кодонного состава гена, а пониженное содержание повторов в генах, вероятно, досталось ему от предка.



**Рис. 26.** Филогенетическое дерево *Mycoplasma*, построенное на основе анализа последовательностей 16S рРНК. Данные взяты из статьи [Peters et al., 2008]. Линией зеленого цвета подчеркнуты виды с параметром  $M > 30$ , а оранжевого – с  $M \leq 30$ .

Как было отмечено выше, *M. pneumoniae* не попала в группу гемоплазм, но у данного организма также наблюдается низкое значение параметра  $M$  и сниженное среднее количество локальных инвертированных повторов на один ген. *M. pneumoniae* является паразитом верхних и нижних дыхательных путей человека. Она может взаимодействовать с поверхностью клеток дыхательного эпителия (реснитчатый эпителий трахей и клетки, выстилающие подслизистые

железы [Collier and Clyde 1971; Powell et al., 1976]) и, возможно, проникать внутрь них [Waites et al., 2008]. Возможно, условия обитания данного организма чем-то схожи с условиями обитания на поверхности эритроцитов, что могло способствовать эволюции данного организма в сторону уменьшения количества повторов (потенциальных вторичных структур) в его генах.

### 3.2.4. Анализ профилей LCI индексов индивидуальных нуклеотидов у *Mycoplasma*

Для более подробного изучения распределения вторичных структур в генах организмов рода *Mycoplasma* были рассчитаны специальные индексы локальной комплементарности для каждого нуклеотида в гене и на его флангах ( $LCI(i, j)$ , где  $i$  – номер гена,  $j$  – номер нуклеотида в гене). После расчетов все гены одного организма выравнивались по старт-кодону (стоп-кодону) трансляции и рассчитывались средние значения индексов  $LCI(i, j)$ . Часть полученных результатов представлена на Рисунках 27-30 (остальные графики в Таблице 5 в Приложении).

У большинства *Mycoplasma* профиль средних значений  $LCI(i, j)$  для 5'- и 3'-районов гена имеет вид, как у *M. fermentans* JER (Рисунки 27-28). В районе старт кодона трансляции наблюдается характерный спад профиля, а в районе стоп кодона – пик. Графики схожи с  $LCI(i, j)$  профилями для других организмов (*E. coli* (Рисунки 18-19), *S. cerevisiae*, *H. sapiens* и др.) [Matushkin et al., 2004].

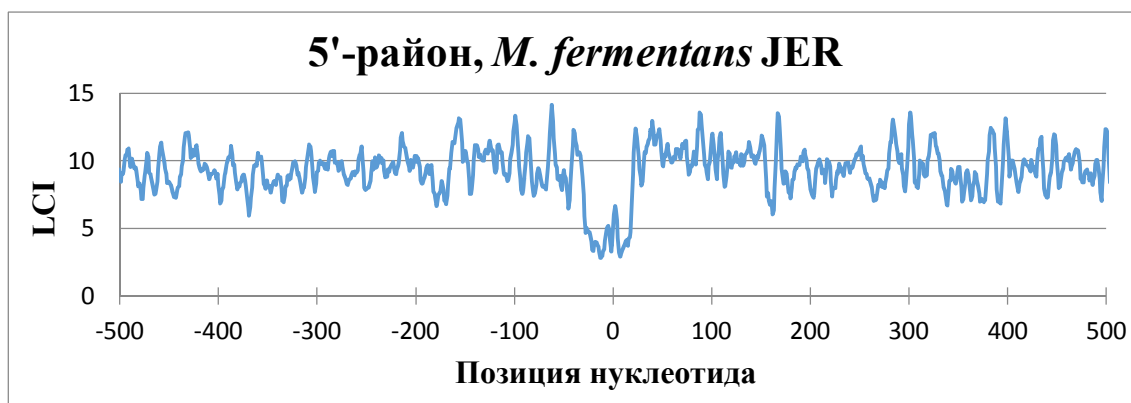


Рис. 27. Профиль средних значений  $LCI(i, j)$  по всем генам *M. fermentans* JER (0 – старт-кодон).

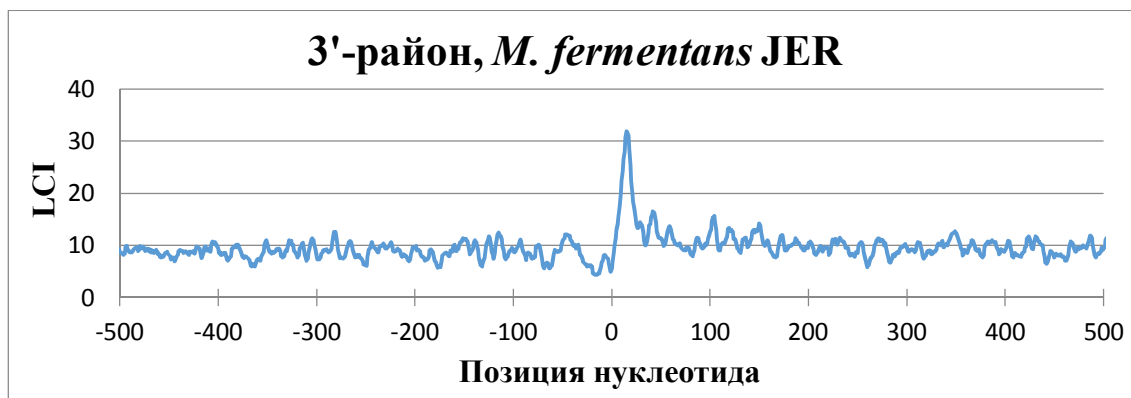


Рис. 28. Профиль средних значений  $LCI(i, j)$  по всем генам *M. fermentans* JER (0 – стоп-кодон).



Спад профиля в 5'-районе, т.е. сниженная стабильность потенциальных шпилек, вероятнее всего, способствует сборке рибосомного комплекса и началу трансляции. Как было указано в разделе 1.2.2.1. главы «Обзор литературы», сниженная стабильность вторичных структур в районе старт-кодона, по-видимому, является универсальной особенностью клеточной жизни. Предполагается, что этот спад стабильности способствует эффективному распознаванию старт-кодона инициаторной тРНК [Gu et al., 2010].

Наличие пика в 3'-районе говорит о повышенной стабильности потенциальных шпилек в данной области, которые могут отвечать, наоборот, за терминацию трансляции или, возможно, транскрипции.

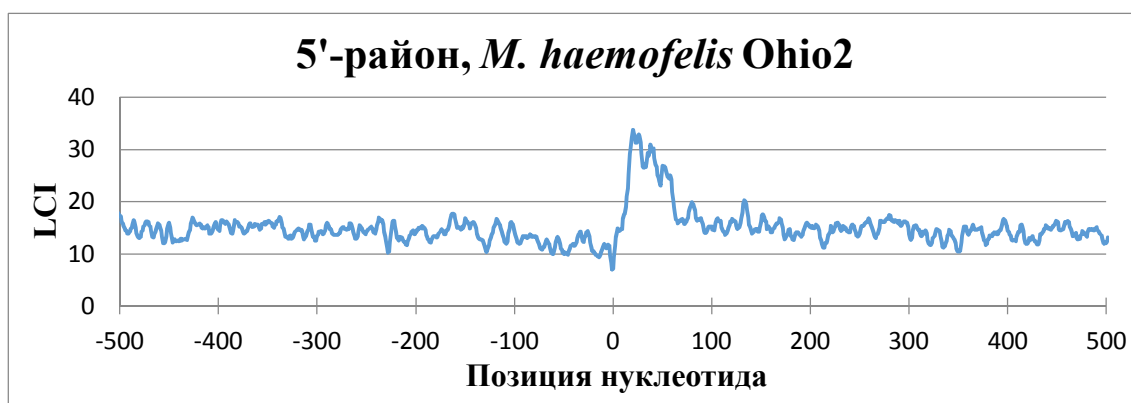


Рис. 29. Профиль средних значений  $LCI(i, j)$  по всем генам *M. haemofelis* Ohio2 (0 – старт-кодон).

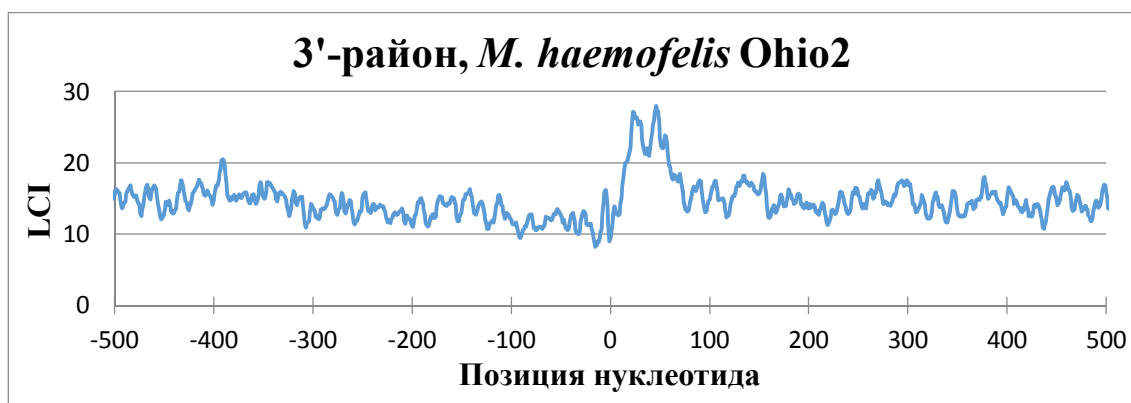


Рис. 30. Профиль средних значений  $LCI(i, j)$  по всем генам *M. haemofelis* Ohio2 (0 – стоп-кодон).

Интересной особенностью обладают профили для 5'-районов у *M. haemofelis*. У данного организма вместо спада профиля наоборот наблюдается его повышение (Рисунок 29), что говорит о повышенной стабильности потенциальных шпилек в данной области. Пока сложно сказать что-либо о причинах данного явления и о том, почему именно *M. haemofelis* обладает данной особенностью. Но, как уже было сказано выше, данный организм обитает в особых условиях (поверхность или внутриклеточное пространство эритроцитов), которые, возможно, способствовали эволюции первичной структуры его генов в сторону уменьшения количества

локальных инвертированных повторов. Если предположить, что скорость элонгации трансляции генов данного организма возросла, возможно, это могло способствовать усилению регуляции на стадии инициации трансляции. Поэтому возможно, что шпильки в 5'-районе могут отвечать именно за регуляцию инициации трансляции у данного штамма. Можно предположить, что у *M. haemofelis* есть особенный механизм регуляции инициации трансляции, отличный от механизмов в других *Mycoplasma*. Для выяснения причин наличия пика в 5'-районе профиля LCI(*i, j*) у *M. haemofelis* необходимы дополнительные исследования, в частности более детальное изучение распределения и строения вторичных структур в данном районе генов этого организма.

### 3.2.5. Подробное исследование профилей LCI индивидуальных нуклеотидов у *Mycoplasma*

В Таблице 5 в Приложении представлены профили средних по всем генам значений LCI индивидуальных нуклеотидов для всех 62 исследованных штаммов микоплазм. При подробном их изучении было установлено, что:

1) у большинства штаммов наблюдается явный спад профиля в районе старт-кодона трансляции (примерно в районе [-20; +20]) и пик за стоп-кодоном длиной, примерно, 40 нуклеотидов (Рисунки 27-28);

2) у *Mycoplasma haemocanis* и *Mycoplasma haemofelis* наблюдается пик как за старт-, так и за стоп-кодоном. Обычный спад профиля в районе старт-кодона не наблюдается (Рисунки 29-30);

3) у *Mycoplasma gallisepticum* наблюдается спад и последующее повышение профиля как в районе старт-, так и стоп-кодона (Рисунки 31-32);

4) у *Mycoplasma pneumoniae* и *Mycoplasma hyopneumoniae* наблюдается спад профиля как в районе старт-, так и в районе стоп-кодона (см. Таблицу 5 в Приложении).

Кроме этого были построены профили LCI индивидуальных нуклеотидов, усредненные не по всем генам организма, а только по:

- а) единственным цистронам в оперонах;
- б) первым цистронам в оперонах;
- в) внутренним цистронам в оперонах;
- г) последним цистронам в оперонах;
- д) 15% высокоэкспрессирующихся генов (с наибольшими значениями EEI);
- е) 15% низкоэкспрессирующихся генов (с наименьшими значениями EEI).

Считалось, что гены, расстояние между которыми меньше или равно 40 нуклеотидам, расположены в одном опероне [Salgado et al., 2000].

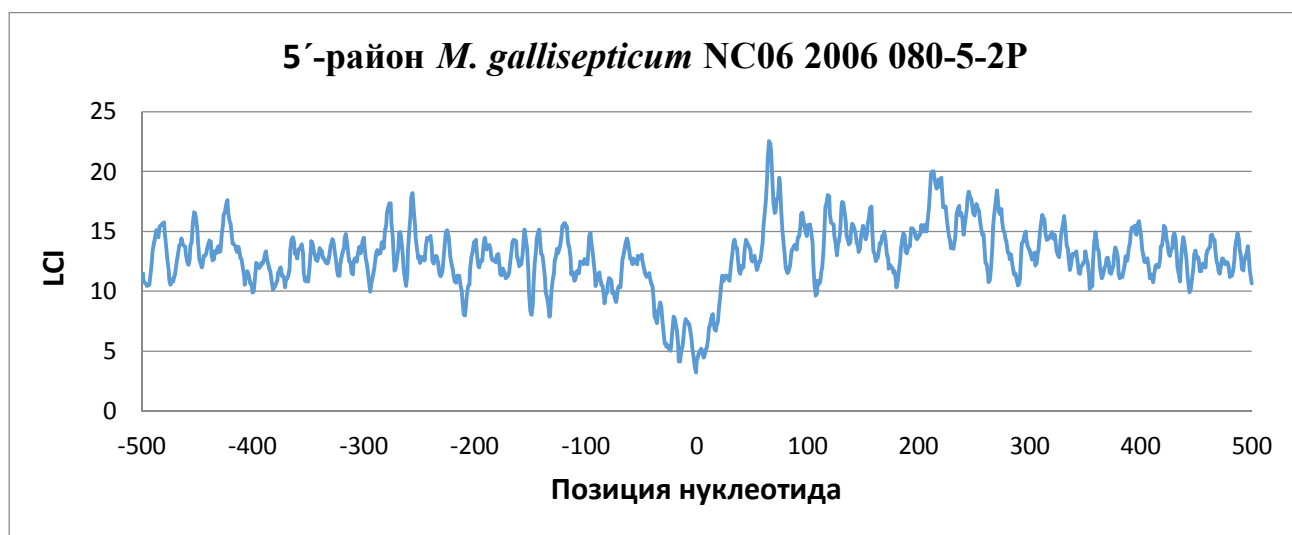


Рис. 31. Профиль средних значений  $LCI(i, j)$  по всем генам *M. gallisepticum* NC06 2006 080-5-2P (0 – старт-кодон).

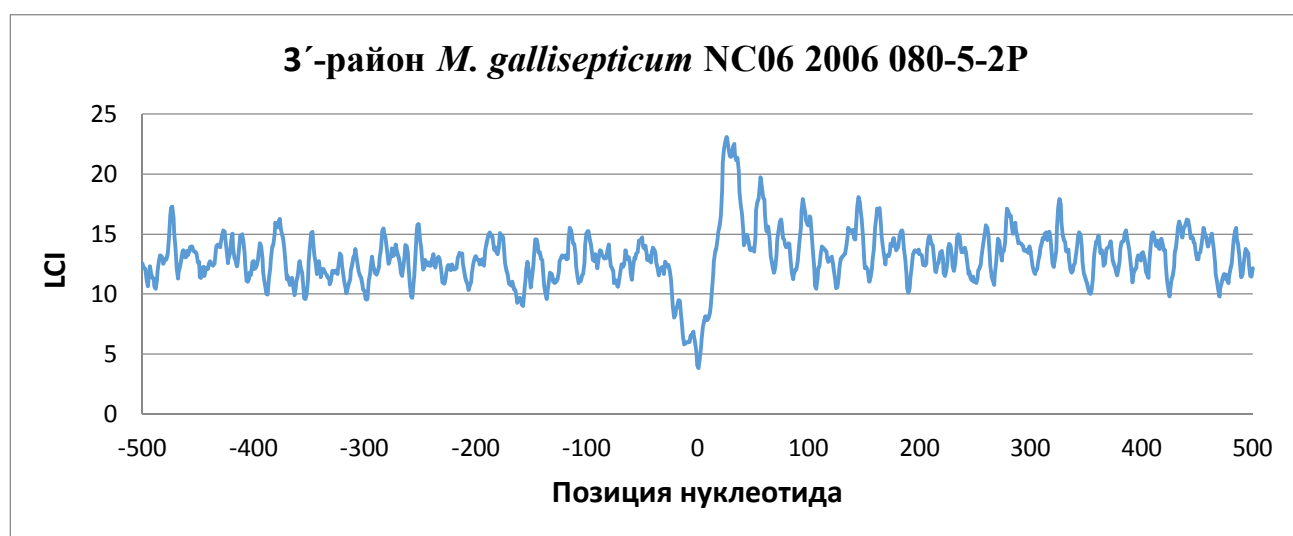
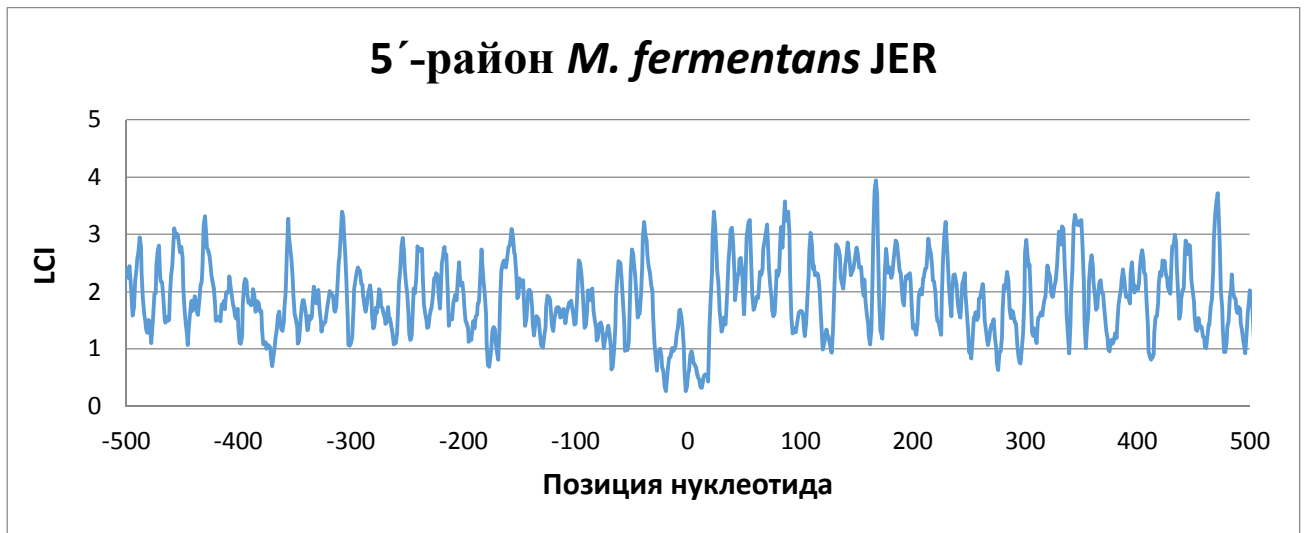


Рис. 32. Профиль средних значений  $LCI(i, j)$  по всем генам *M. gallisepticum* NC06 2006 080-5-2P (0 – стоп-кодон).

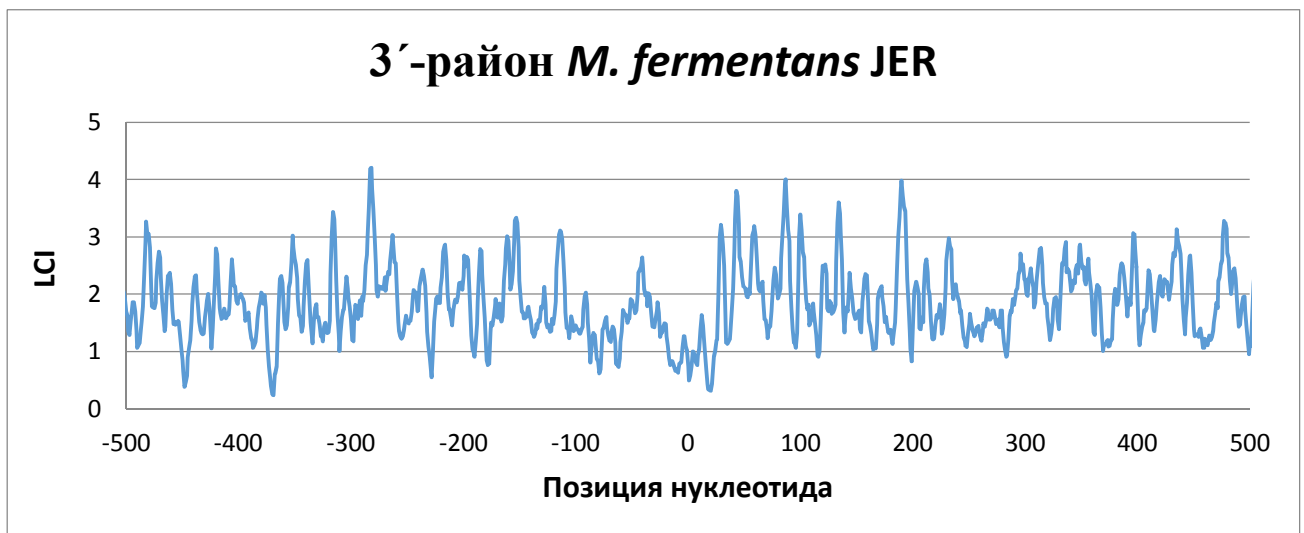
Анализ этих данных показал, что при рассмотрении только внутренних цистронов, практически у всех штаммов наблюдается спад как в районе старт-, так и стоп-кодона (Рисунки 33-34). Явных пиков в районе стоп-кодона не наблюдается.

Исключением являются штаммы *Mycoplasma haemofelis* Ohio2 и *Mycoplasma haemofelis* Langford1. У них наблюдается такие же пики в районах старт- и стоп-кодонов (Рисунки 35-36), как и на  $LCI(i, j)$  профилях, усредненных по всем генам (Рисунки 29-30).

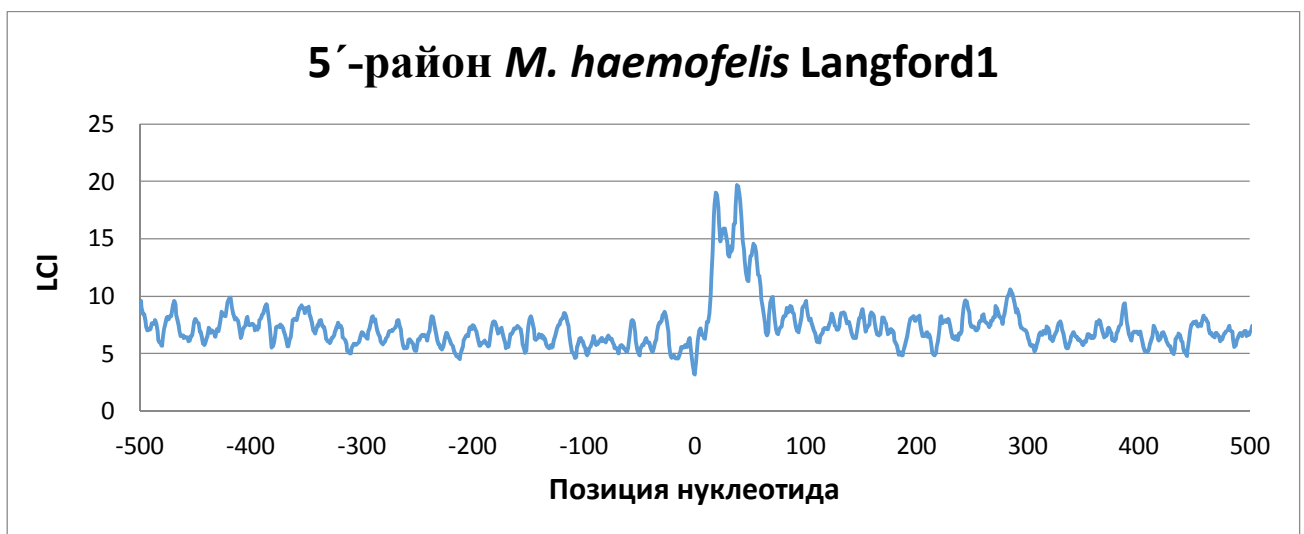
Обнаруженные различия в профилях средних значений  $LCI(i, j)$  у *Mycoplasma* говорят о наличии локальных особенностей нуклеотидных последовательностей в зависимости от их положения (начало/конец гена, первый/внутренний/последний цистрон в опероне). Дальнейшее более подробное изучение данных особенностей может помочь лучше понять механизмы регуляции трансляции генов различных организмов.



**Рис. 33.** Профиль средних значений  $LCI(i, j)$  по внутренним цистронам в оперонах *M. fermentans* JER (0 – старт-кодон).



**Рис. 34.** Профиль средних значений  $LCI(i, j)$  по внутренним цистронам в оперонах *M. fermentans* JER (0 – стоп-кодон).



**Рис. 35.** Профиль средних значений  $LCI(i, j)$  по внутренним цистронам в оперонах *M. haemofelis* Langford1 (0 – старт-кодон).

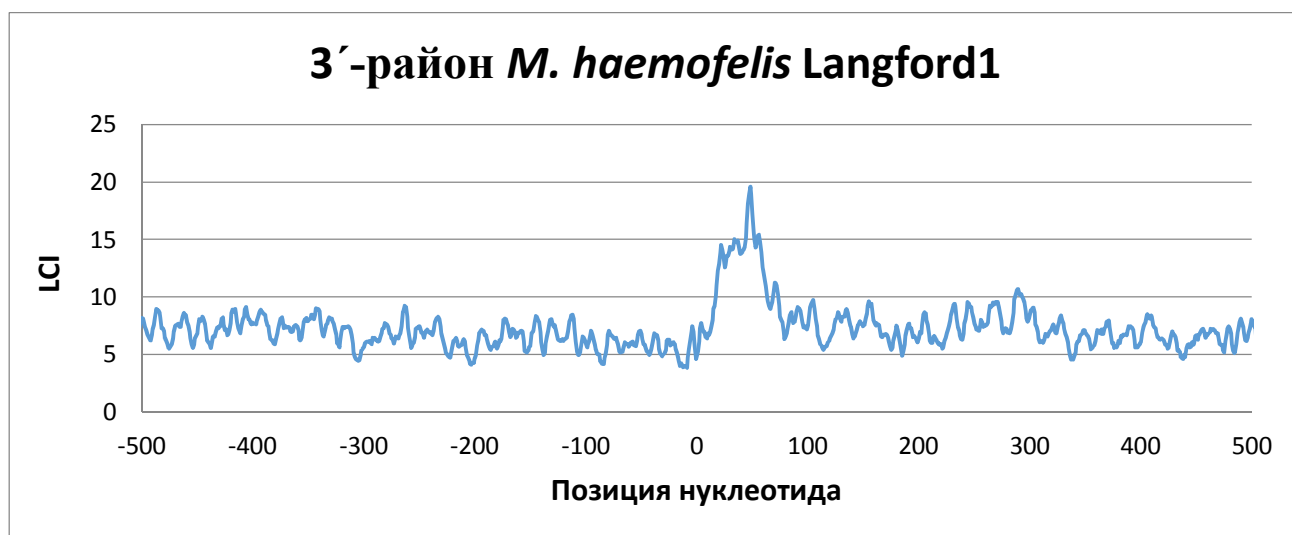


Рис. 36. Профиль средних значений  $LCI(i, j)$  по внутренним цистронам в оперонах *M. haemofelis* Langford1 (0 – стоп-кодон).

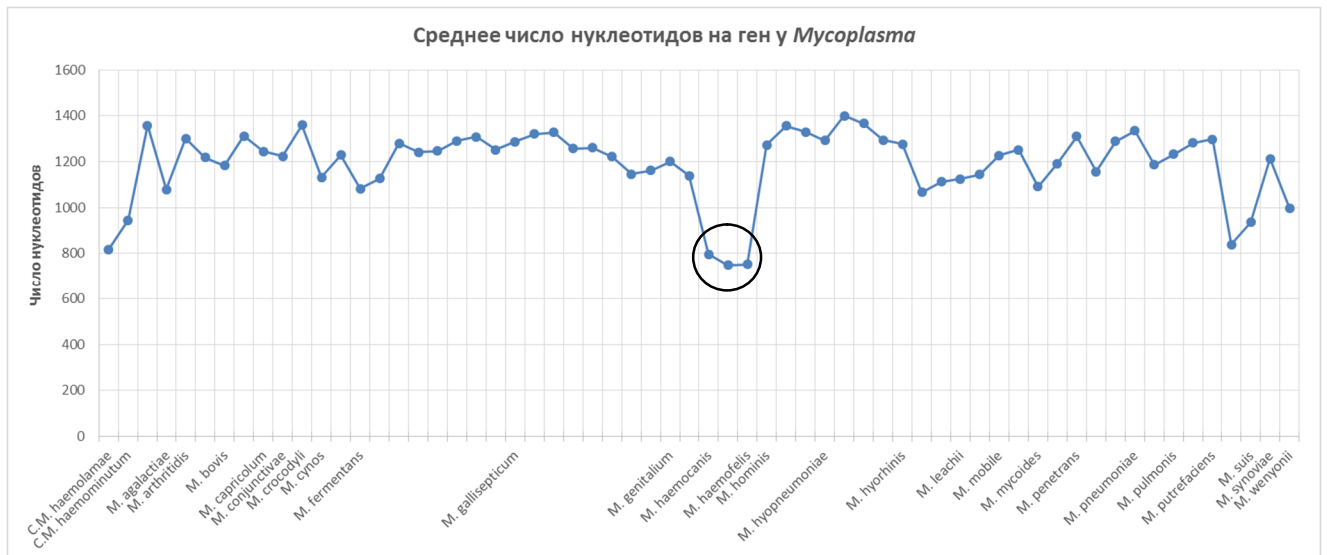
### 3.2.6. Исследование оперонной структуры генов различных штаммов *Mycoplasma*

При изучении оперонной структуры генов исследуемых *Mycoplasma* была обнаружена следующая особенность. У «особенных» видов *M. haemocanis* и *M. haemofelis* (Рисунок 37, черный овал) количество генов примерно в 1,5-2 раза больше, чем у остальных микоплазм. Также показано, что этот избыток получается за счет увеличения количества именно внутренних цистронов в оперонах, а не единственных, как у всех остальных. Таким образом, за счет того, что у данных организмов количество генов выше среднего значения в 2 раза, а размер генома только в 1,28, гены в нем расположены плотнее друг к другу (на расстоянии  $\leq 40$  нуклеотидов) (Рисунок 38, черный круг). Это может приводить к формированию длинных оперонов, состоящих из большого числа цистронов.

Как было показано в разделе 1.1. главы «Обзор литературы» у прокариот, в том числе и микоплазм [Uenoyma and Miyata, 2005; Waldo and Krause, 2006], известно такое явление как реинициация трансляции. Расположение большого количества генов близко друг к другу у *M. haemofelis* может приводить к нежелательной реинициации трансляции генов, расположенных после гена, действительно требующего трансляции. Поэтому для каждого гена в таких больших оперонах требуется более эффективная терминация трансляции. Этому могут способствовать более стабильные вторичные структуры в 5' и 3' районах. Как было показано на Рисунках 35-36 в предыдущем разделе, внутренние цистроны в оперонах *M. haemofelis* характеризуются более стабильными вторичными структурами в 5' и 3' районах мРНК, в отличие от остальных микоплазм. Таким образом, можно предположить, что эти вторичные структуры (пики на Рисунках 35-36) уменьшают вероятность реинициации трансляции.



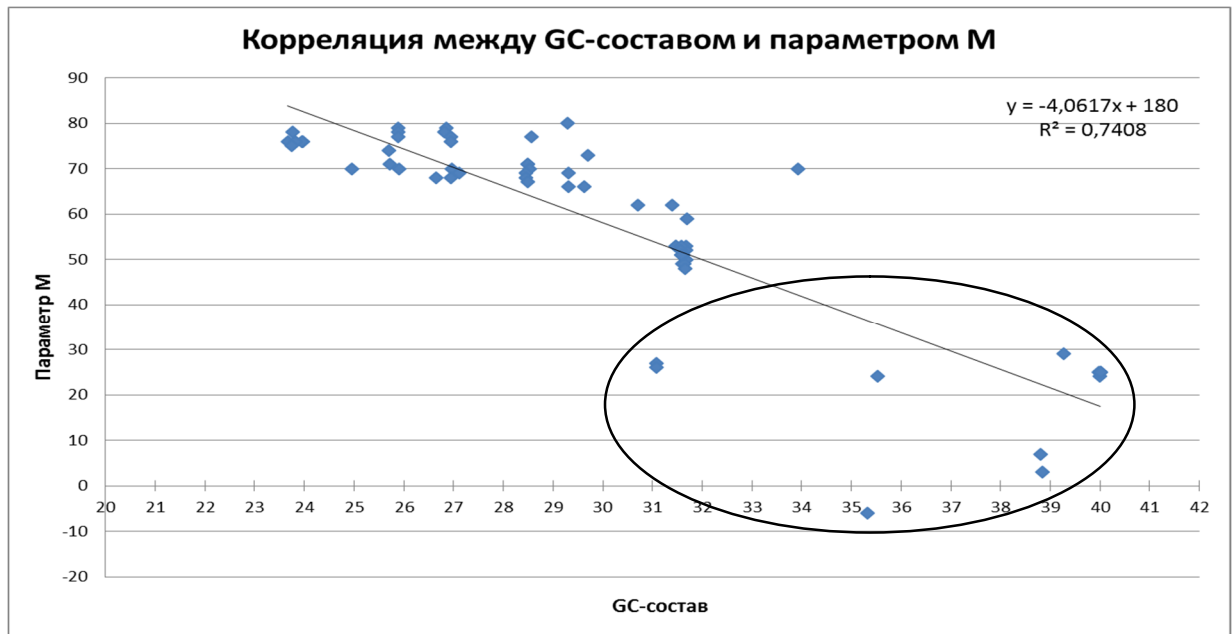
**Рис. 37.** Количество генов каждого типа в разных штаммах *Mycoplasma*. Черным овалом выделены штаммы *M. haemocanis* str. Illinois, *M. haemofelis* str. Langford1 и *M. haemofelis* str. Ohio2.



**Рис. 38.** Среднее число нуклеотидов на один ген в разных штаммах *Mycoplasma*. Черным кругом выделены штаммы *M. haemocanis* str. Illinois, *M. haemofelis* str. Langford1 и *M. haemofelis* str. Ohio2.

### 3.2.7. Связь между GC-составом и эволюционной оптимизацией первичной структуры генов *Mycoplasma* для повышения эффективности элонгации трансляции

Как было показано в главе «Обзор литературы», эффективность трансляции может зависеть от GC-состава генов. Для проверки данного факта для 62 исследуемых штаммов *Mycoplasma* был построен график зависимости параметра  $M$ , показывающего эффективность определения ГРБ как высокоэкспрессируемых, от GC-состава генома данных организмов (Рисунок 39).



**Рис. 39.** Зависимость параметра М от GC-состава генома для 62 штаммов *Mycoplasma* ( $r = -0,85$ ;  $P < 1,97 \times 10^{-18}$ ). Большинство штаммов с параметром М < 30 (выделены черным овалом) характеризуются наибольшими значениями GC-состава генома.

Как видно из Рисунка 39, имеет место достоверная отрицательная корреляция. Коэффициент корреляции Спирмена  $r = -0,85$  ( $P < 1,97 \times 10^{-18}$ ) между GC-составом генома и степенью эволюционной оптимизации первичной структуры его генов для повышения эффективности элонгации трансляции. Отсюда следует, что большинство штаммов с параметром  $M \leq 30$  характеризуются наибольшими значениями GC-состава генома. Этот факт хорошо согласуется с предположением о том, что у данных «особенных» микоплазм локальные инвертированные повторы редко встречаются в геноме из-за повышенной стабильности шпилек, которые они могут сформировать. Так как GC-богатые шпильки более стабильны, чем АТ-богатые (см. раздел 3.2.2. главы «Результаты и обсуждение»).

Также для исследованных микоплазм были рассчитаны коэффициенты корреляции Спирмена между GC-составом генов и их ЕЕI. Дисперсионный анализ данных коэффициентов корреляции показал достоверное различие между группами организмов, принадлежащих к разным типам по эволюционной оптимизации первичной структуры их генов (ЕЕI1-ЕЕI4). Результаты представлены на Рисунке 40. К первой и третьей группам относятся всего по два организма (см. Рисунок 24), поэтому о достоверных отличиях и каких-либо выводах говорить сложно. Среднее значение коэффициентов корреляции для микоплазм из 2-ой группы равно 0,44, а из 4-ой – 0,83. Разница между ними достоверна ( $P < 1,52 \times 10^{-14}$ ).

Таким образом, у организмов, относящихся ко второй группе (ЕЕI2) по типу эволюционной оптимизации, у которых эффективность элонгации трансляции зависит только от числа локальных инвертированных повторов в генах, корреляция между GC-составом гена и

EEI выше, чем у организмов из 4-ой группы, у которых эффективность элонгации трансляции зависит как от числа повторов, так и от кодонного состава. Высокое положительное значение коэффициента корреляции 0,83 между GC-составом и EEI говорит о том, что в генах с высоким GC-составом снижено число локальных инвертированных повторов, которые потенциально могут образовать стабильные вторичные структуры в мРНК. Более низкий коэффициент корреляции 0,44 между GC-составом и EEI для организмов из 4-ой группы, видимо, связан с тем, что учитывается не только количество повторов, но и кодонный состав генов.

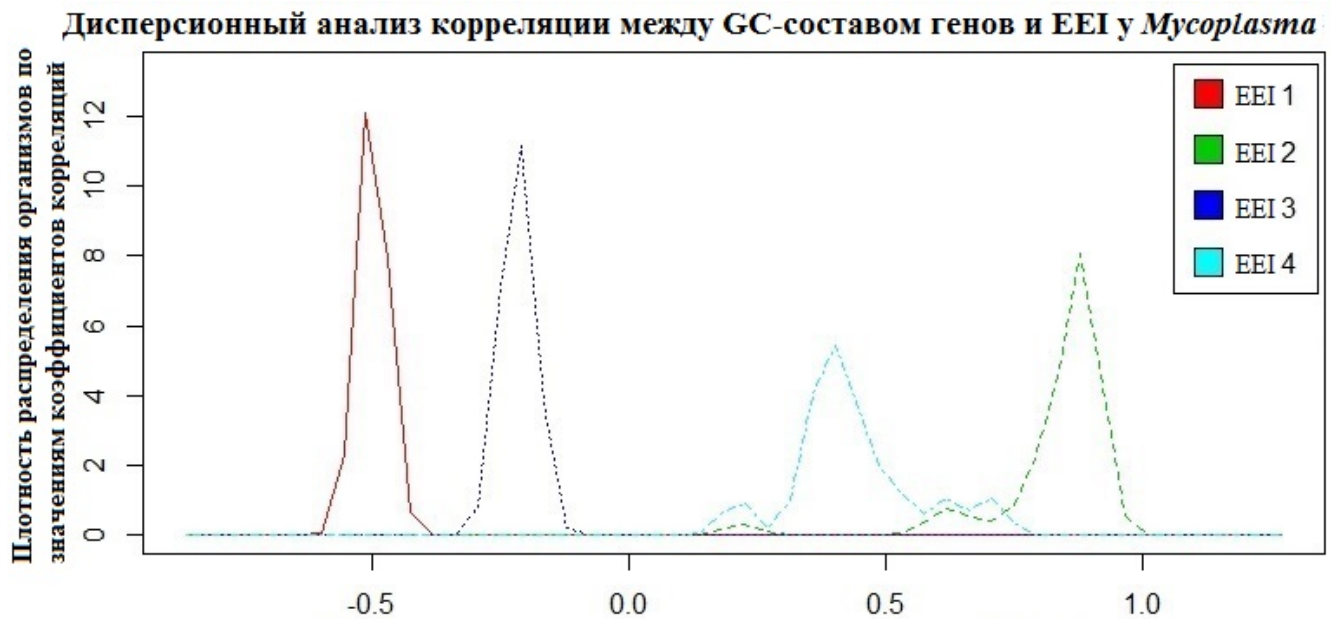


Рис. 40. Коэффициент корреляции (ось абсцисс) между GC-составом гена и EEI у *Mycoplasma*.

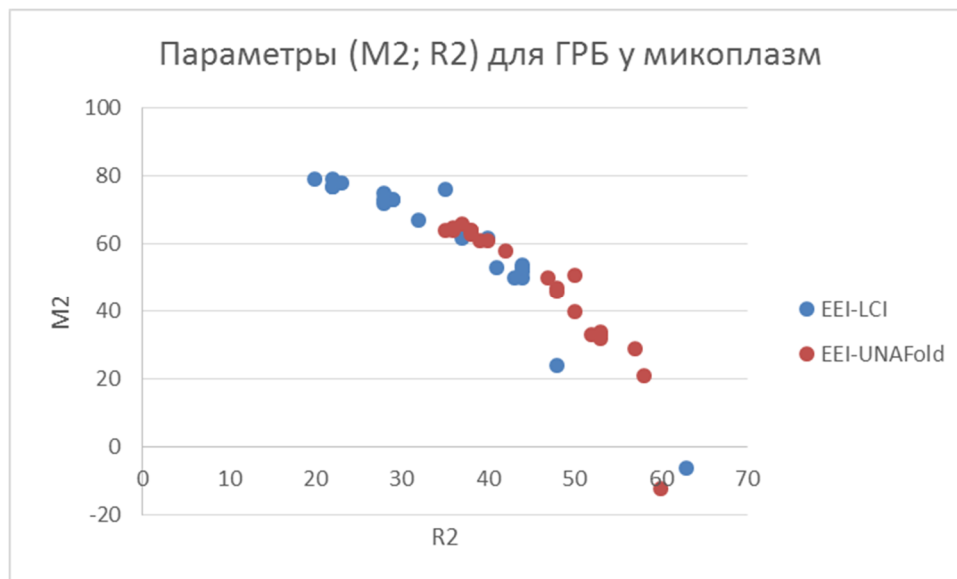
### 3.2.8. Использование программы UNAFold для предсказания вторичной структуры мРНК у *Mycoplasma*

Поскольку подсчет повторов в генах может недостаточно точно отражать количество шпилек в мРНК, для этой задачи была использована специальная программа UNAFold, позволяющая рассчитывать вторичную структуру нуклеотидных последовательностей [Markham and Zuker, 2008]. В качестве объектов исследования были выбраны микоплазмы, так как у большинства из них работает второй тип EEI (Рисунок 24, раздела 3.2.1. главы «Результаты и обсуждение»), т.е. на эффективность элонгации трансляции оказывает влияние в основном количество локальных инвертированных повторов в ОРС (потенциальных вторичных структур в мРНК).

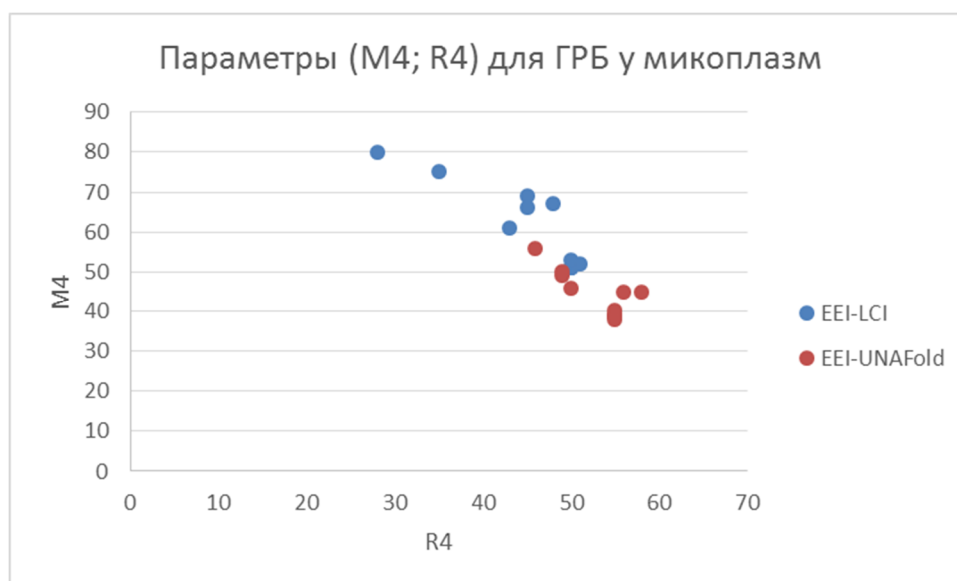
Суть исследования заключалась в смене метода учета шпилек в мРНК с простого подсчета локальных инвертированных повторов (LCI1 и LCI2) на непосредственный фолдинг участков мРНК гена определенной длины при помощи программы UNAFold и подсчете



образовавшихся шпилек и их энергии. Далее сравнивалось, какая из версий программы EEI-LCI или EEI-UNAFold лучше определяет ГРБ микоплазм как высокоэкспрессируемые.



**Рис. 41.** Сравнение значений параметров (M2; R2) для ГРБ у микоплазм для программ EEI-LCI и EEI-UNAFold.



**Рис. 42.** Сравнение значений параметров (M4; R4) для ГРБ у микоплазм для программ EEI-LCI и EEI-UNAFold.

На Рисунке 41 показаны значения пар параметров (M2; R2) для микоплазм, у которых работает второй тип EEI, для каждой из программ. Среднее значения M2 достоверно выше ( $P < 2,06 \times 10^{-2}$ ), а R2 достоверно ниже ( $P < 5,11 \times 10^{-4}$ ) для EEI-LCI. Аналогичные результаты, но для пар (M4; R4) показаны на Рисунке 42. Среднее значения M4 достоверно выше ( $P < 4,79 \times 10^{-4}$ ), а R2 достоверно ниже ( $P < 8,22 \times 10^{-3}$ ) для EEI-LCI. Таким образом, EEI-LCI лучше, чем EEI-UNAFold, определяет ГРБ у микоплазм как высокоэкспрессируемые.

Причины, по которым учет локальных инвертированных повторов позволяет лучше определять ГРБ как высокоэкспрессируемые, чем непосредственный подсчет шпилек в мРНК при помощи программы UNAFold, до конца не ясны. Этот вопрос требует дополнительного подробного исследования. Однако стоит заметить, что, если некоторый набор параметров позволит программе UNAFold так же хорошо определять ГРБ как высокоэкспрессируемые, как это делает LCI, последний все-равно будет значительно выигрывать по скорости расчетов.

### **3.3. Исследование оптимизации первичной структуры генов архей в процессе эволюции**

В рамках работы по изучению эволюции белков представителей царства Архей был проведен анализ нуклеотидных последовательностей генов, кодирующих исследуемые белки, на предмет значительных изменений в уровне эффективности/скорости элонгации трансляции в процессе эволюции.

Проведенный ранее анализ царства Архей выявил, что у большинства его представителей эффективность элонгации трансляции генов зависит от распределения в них локальных инвертированных повторов, т.к. в EEI2 и EEI4 учитываются локальные инвертированные повторы (Рисунок 22, раздел 3.1.2. главы «Результаты и обсуждение»).

В ходе работы был проведен анализ нуклеотидных последовательностей генов как ныне существующих организмов, так и их предков, чьи последовательности были специально реконструированы. Список проанализированных таксономических групп архей и виды, принадлежащие этим группам приведены в Таблице 6 в Приложении. Отсутствие полных геномов предковых организмов не позволяет рассчитывать индексы, учитывающие кодонный состав (EEI1, EEI4 и EEI5). Кроме этого, поскольку реконструированные последовательности имеют в определенной степени вероятностный характер, их анализ дает лишь приближенные значения индексов EEI. Поэтому в качестве анализируемого параметра был выбран тот тип индекса EEI, который менее чувствителен к случайным синонимичным заменам кодонов.

#### **3.3.1. Исследование влияния синонимичных замен в генах организма на значения индекса EEI**

Для выяснения, какой тип индекса EEI менее других чувствителен к случайным синонимичным заменам (мутациям) кодонов, было проведено дополнительное исследование. Результаты показали, что для высокоэкспрессирующихся генов современных организмов значение индекса EEI2 изменяется меньше всего по сравнению с остальными типами EEI при случайных заменах кодонов гена на синонимичные.

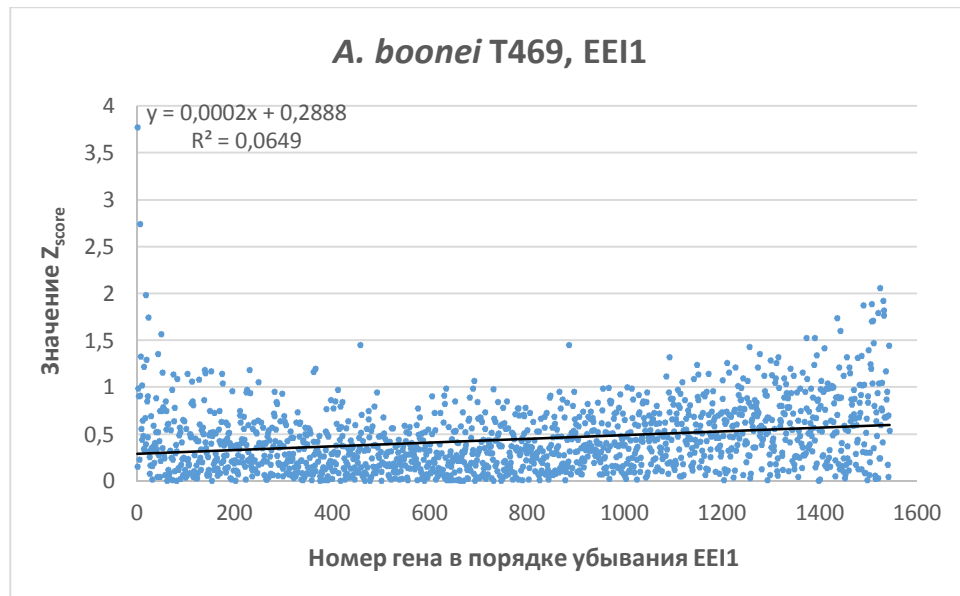
Исследование проводилось следующим образом. Первоначальная программа EtoE была модифицирована так, чтобы для каждого гена организма можно было провести цикл случайных замен кодонов на синонимичные и заново рассчитать индексы EEI. Т.е. каждый кодон в нуклеотидной последовательности с заранее заданной вероятностью заменялся на один из синонимичных. Выбор одного кодона из группы синонимичных происходил случайным образом с равной вероятностью для каждого кодона. Подобный расчет был проведен для всех генов *Escherichia coli* и *Aciduliprofundum boonei*. Эти организмы были выбраны как характерные представители групп организмов, в которых работает первый и второй тип EEI соответственно. *A. boonei* к тому же принадлежит к царству Архей.

Первоначальная последовательность каждого гена «мутировалась» 10000 раз (на каждой итерации цикла мутации бралась исходная последовательность гена). Вероятность замены определенного кодона на произвольный синонимичный кодон составляла 10-50% (в зависимости от эксперимента). По полученным новым значениям EEI рассчитывалось среднее и стандартное отклонение. Далее для каждого гена рассчитывалось значение  $Z_{score}$  равное отношению модуля разности первоначального значения  $EEI_0$  до мутаций и среднего значения  $EEI_{cp}$  после циклов мутаций к стандартному отклонению:

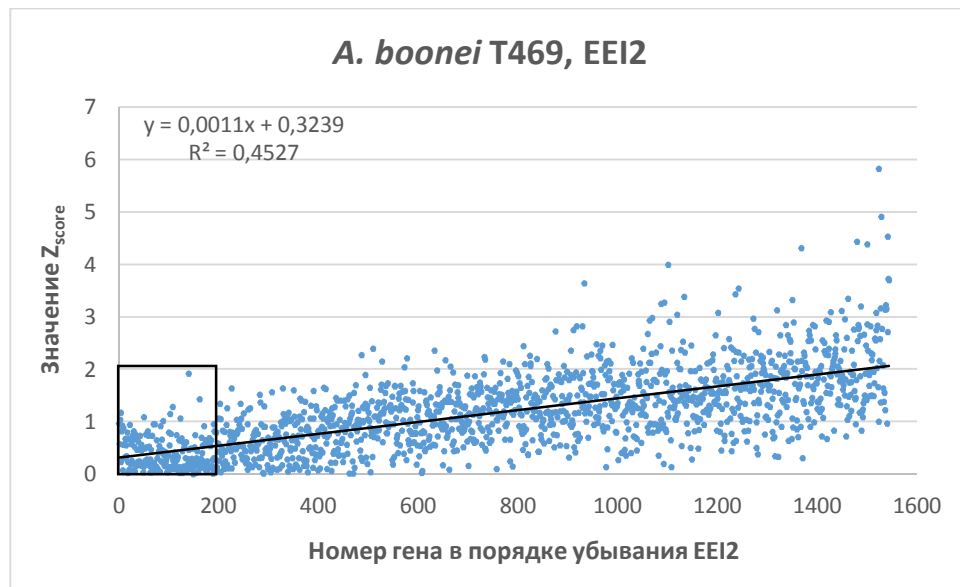
$$Z_{score} = \frac{|EEI_0 - EEI_{cp}|}{\sigma}, \quad (28) \quad EEI_{cp} = \sum_{i=1}^N EEI_i / N, \quad (29)$$

где  $EEI_0$  – значение индекса EEI исходной последовательности гена,  $EEI_i$  – значение индекса EEI гена после  $i$ -ой мутации,  $EEI_{cp}$  и  $\sigma$  – среднее и стандартное отклонение от среднего соответственно,  $N$  – количество циклов мутаций.

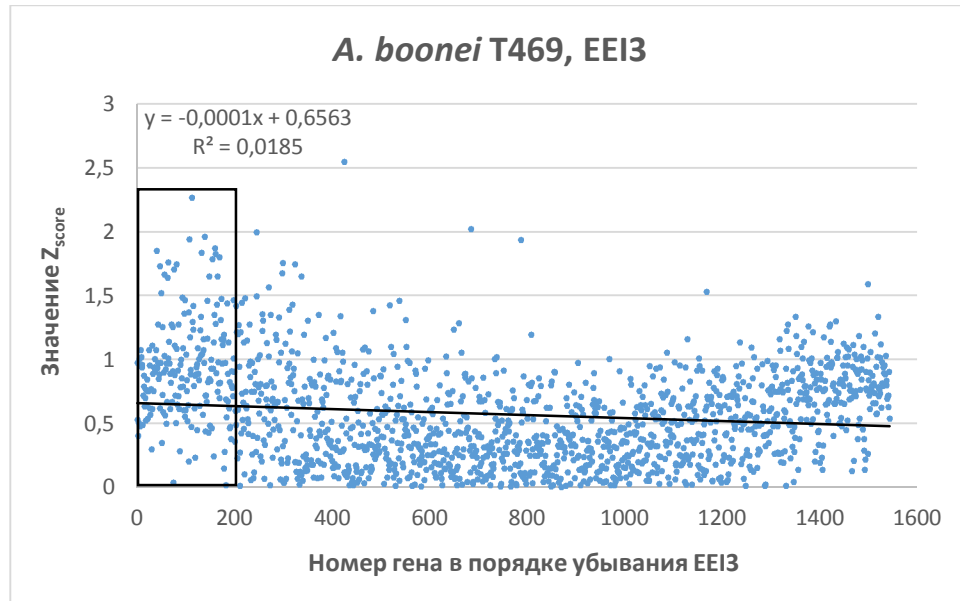
Полученные значения для всех генов *A. boonei* были нанесены на графики. Приведены данные для следующих значений параметров мутирования: 10000 циклов с вероятностью мутации 25% (Рисунки 43-47). Аналогичные результаты были получены при других значениях параметров. Как видно из графиков, значения  $EEI_2$  для генов с высокой эффективностью элонгации трансляции изменяются меньше под воздействием мутаций, чем значения  $EEI_3$ . Критерий Манна-Уитни показал достоверное различие ( $P < 2,2 \times 10^{-18}$ ) между средними значениями  $Z_{score}$  для выборок из первых 200 генов с наибольшими  $EEI_2$  и  $EEI_3$  (Рисунки 44-45). Для сравнения приведены графики для *E. coli* для индексов  $EEI_2$  и  $EEI_3$  (Рисунки 48-49).



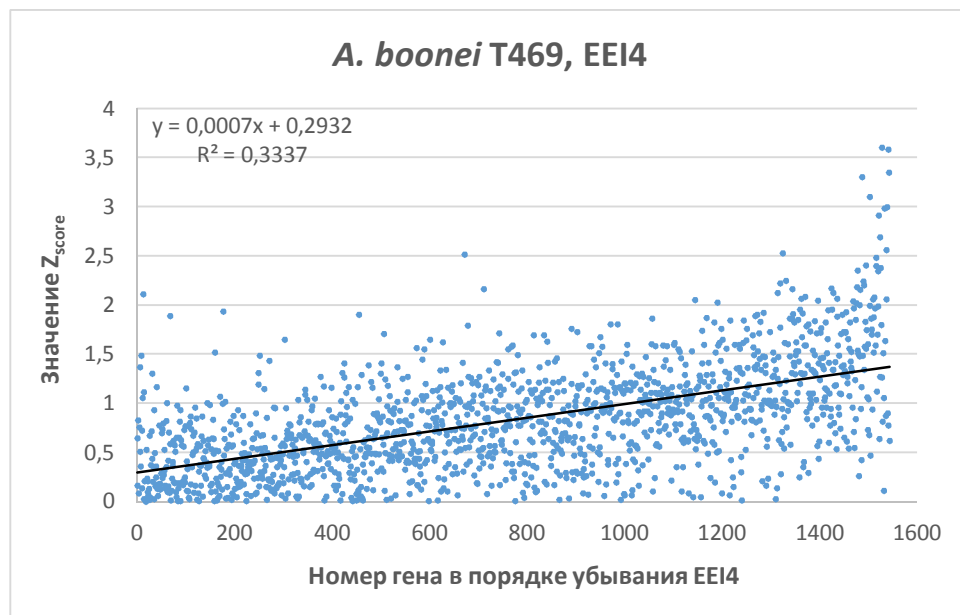
**Рис. 43.** Значения параметра  $Z_{score}$  для индекса EEI1 генов *A. boonei* T469 после «мутирования» ( $r = 0,25$ ;  $P < 2,73 \times 10^{-24}$ ). 10000 циклов мутаций с вероятностью мутации 25%.



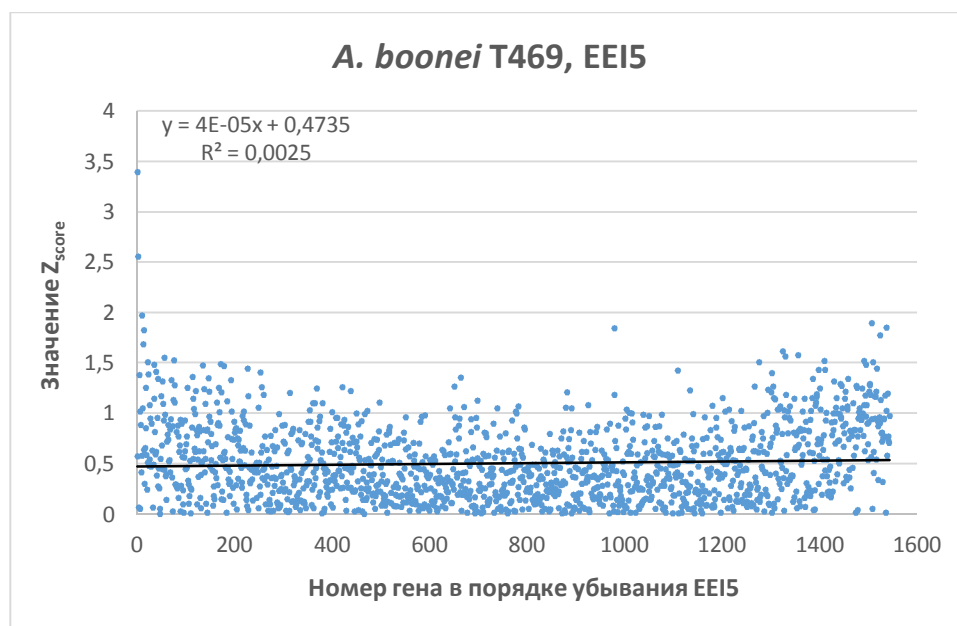
**Рис. 44.** Значения параметра  $Z_{score}$  для индекса EEI2 генов *A. boonei* T469 после «мутирования» ( $r = 0,67$ ;  $P < 4,5 \times 10^{-204}$ ). 10000 циклов мутаций с вероятностью мутации 25%. Черным прямоугольником выделены первые 200 генов для сравнения с EEI3.



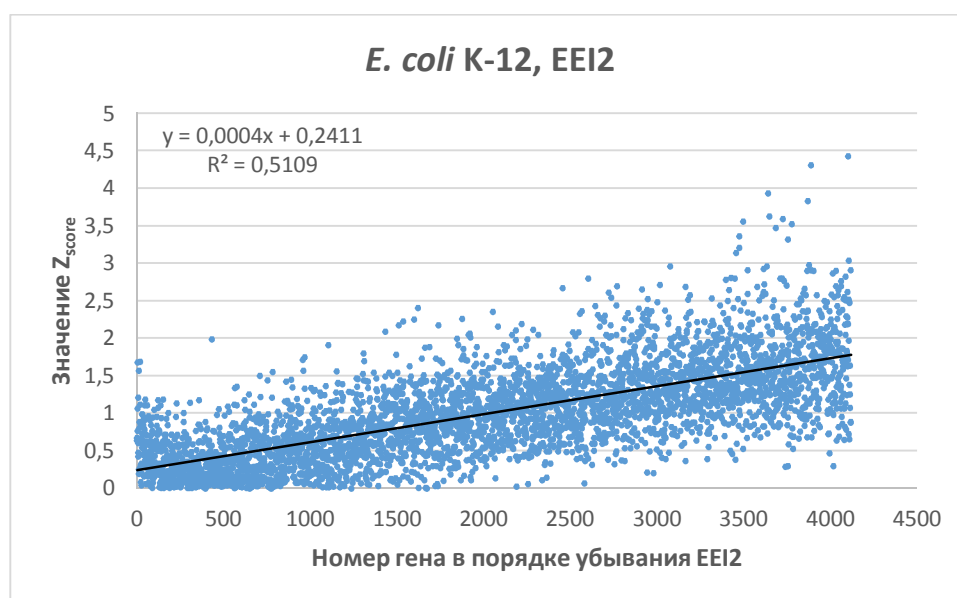
**Рис. 45.** Значения параметра  $Z_{score}$  для индекса EEI3 генов *A. boonei* T469 после «мутирования» ( $r = -0,13$ ;  $P < 8,36 \times 10^{-8}$ ). 10000 циклов мутаций с вероятностью мутации 25%. Черным прямоугольником выделены первые 200 генов для сравнения с EEI2.



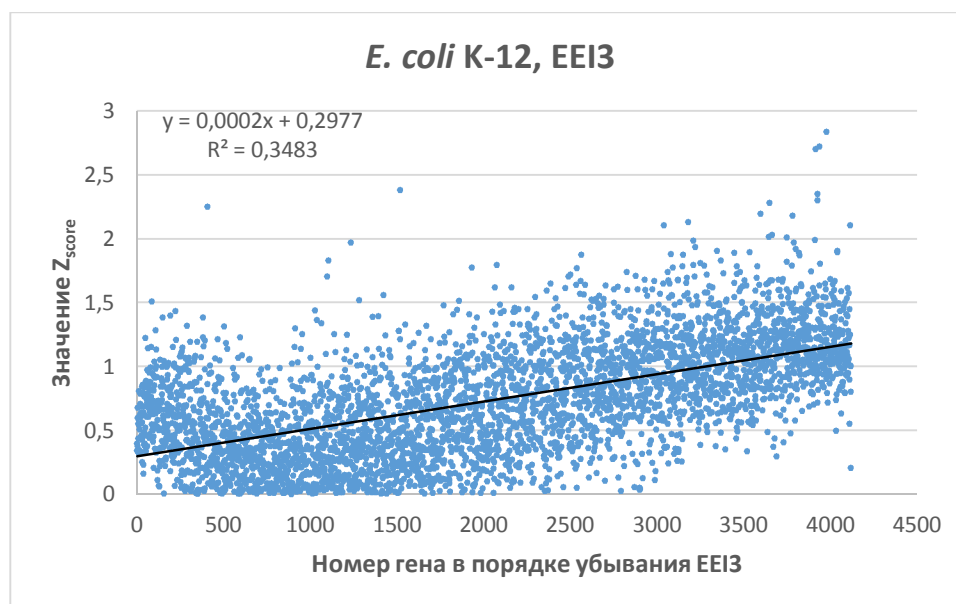
**Рис. 46.** Значения параметра  $Z_{score}$  для индекса EEI4 генов *A. boonei* T469 после «мутирования» ( $r = 0,58$ ;  $P < 3,8 \times 10^{-138}$ ). 10000 циклов мутаций с вероятностью мутации 25%.



**Рис. 47.** Значения параметра  $Z_{score}$  для индекса EEI5 генов *A. boonei* T469 после «мутирования» ( $r = 0,05$ ;  $P < 5,1 \times 10^{-2}$ ). 10000 циклов мутаций с вероятностью мутации 25%.



**Рис. 48.** Значения параметра  $Z_{score}$  для индекса EEI2 генов *E. coli* K-12 после «мутирования» ( $r = 0,71$ ;  $P < 1,3 \times 10^{-204}$ ). 10000 циклов мутаций с вероятностью мутации 25%.



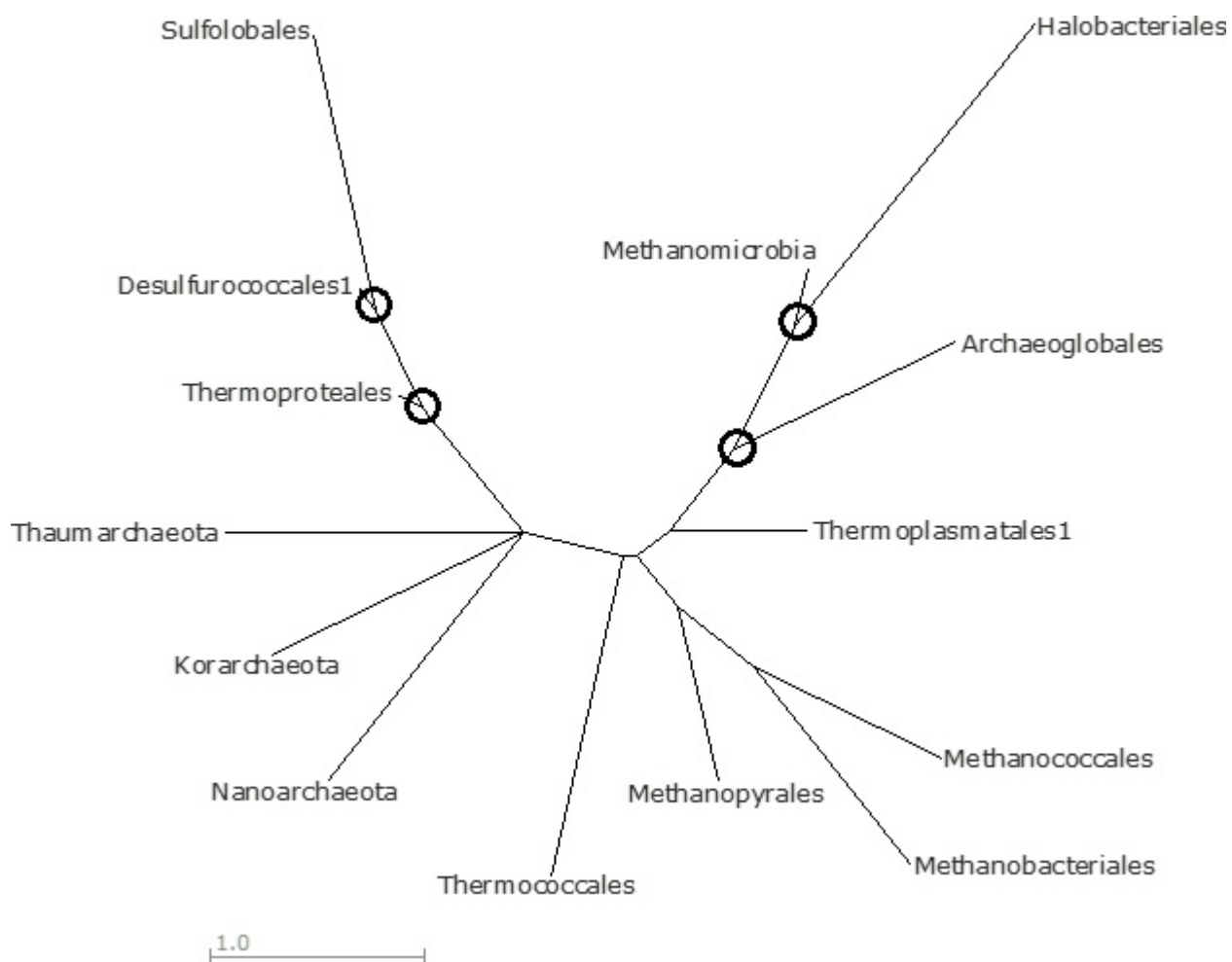
**Рис. 49.** Значения параметра  $Z_{score}$  для индекса EEI3 генов *E. coli* K-12 после «мутирования» ( $r = 0,59$ ;  $P < 1,3 \times 10^{-204}$ ). 10000 циклов мутаций с вероятностью мутации 25%.

Таким образом, значения EEI2 для генов с высокой эффективностью элонгации трансляции изменяются меньше под воздействием мутаций, чем значения EEI3. С учетом отсутствия возможности рассчитывать индексы EEI1, EEI4 и EEI5 у предковых последовательностей для анализа был выбран индекс EEI2.

### 3.3.2. Анализ предковых последовательностей генов архей

Для анализа брались нуклеотидные последовательности генов ныне существующих организмов, принадлежащих к царству Архей, а также нуклеотидные последовательности генов их предков. Предковые последовательности генов были реконструированы К. В. Гунбиным на основе трех альтернативных предковых последовательностей белков (реконструированных с использованием WAG, LG и JTT матриц при помощи пакета RAxML v.7.4.2 и пакета prank v.121218 для реконструкции паттерна делеций/инсерций), а также с использованием информации о кодирующих современные белки последовательностях ДНК. Реконструкция производилась на основе принципа парсимонии – минимального количества нуклеотидных замен, требующихся для объяснения замен аминокислот.

После того, как последовательности генов были реконструированы, для каждой из них рассчитывался индекс EEI2. Результаты записывались в соответствующие им позиции на построенных в процессе реконструкции филогенетических деревьях. Далее проводился анализ данных деревьев для обнаружения наибольших изменений в значениях индекса EEI2 в соседних узлах. На Рисунке 49 приведено консенсусное дерево с основными узлами.



**Рис. 50.** Консенсусное дерево Archaea. Черными кружками отмечены соседства, в которых происходили наиболее значимые изменения в значениях EEI2 (см. текст).

Использовались два способа оценки эволюционных изменений значений индекса EEI2. Первый способ основан на учете значений индекса только самих последовательностей. Для этого для каждого соседства рассчитывалось значение  $Z_{score}$  по формуле аналогичной Формуле 28. Второй способ учитывал кроме значений индексов самих последовательностей еще и значения их ближайших предков. Для этого вместо исходного индекса EEI2 использовалась величина равная модулю разности  $|EEI2_{предок} - EEI2_{потомок}|$ . Далее для данной величины аналогично первому способу вычислялся  $Z_{score}$ . Итоговые результаты в виде основных соседств филогенетического дерева архей с наибольшими значениями  $Z_{score}$  (Рисунок 50, черные круги) представлены в Таблице 7. Также в таблице приведены условия обитания организмов, принадлежащих к данным группам. Остальные соседства, не представленные в таблице, характеризовались значительно более низкими значениями  $Z_{score}$ , что не позволяет говорить о достоверных различиях между ними.



Таблица 7. Основные соседства на филогенетическом дереве, в которых происходят наиболее существенные изменения эффективности элонгации трансляции генов, измеренной с помощью EEI2.

Соседство	Значение $Z_{score}$	Изменение условий обитания в соседстве
<b>Учет значения индекса только самой последовательности</b>		
<b>Halobacteriales-Methanomicrobia</b>	2,13	<b>Halobacteriales:</b> оптимум $T=31-50^{\circ}\text{C}$ ; питание органическим субстратом; аэробы; галофилы
		<b>Methanomicrobia:</b> оптимум $T=0-70^{\circ}\text{C}$ ; питание неорганическим субстратом; анаэробы; галофобы
<b>Desulfurococcales1-Thermoproteales</b>	1,79	<b>Desulfurococcales1:</b> оптимум $T=85-106^{\circ}\text{C}$ ; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности и любых глубинах
		<b>Thermoproteales:</b> оптимум $T=75-104^{\circ}\text{C}$ ; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности
<b>Desulfurococcales1-Sulfolobales</b>	1,55	<b>Desulfurococcales1:</b> оптимум $T=85-106^{\circ}\text{C}$ ; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности и любых глубинах
		<b>Sulfolobales:</b> оптимум $\text{pH}\sim 2$ ; оптимум $T=40-85^{\circ}\text{C}$ ; питание органическим и неорганическим субстратом; факультативные аэробы
<b>Учет значения индекса как самой последовательности, так и ближайшего предка</b>		
<b>Desulfurococcales1-Thermoproteales</b>	2,16	<b>Desulfurococcales1:</b> оптимум $T=85-106^{\circ}\text{C}$ ; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности и любых глубинах
		<b>Thermoproteales:</b> оптимум $T=75-104^{\circ}\text{C}$ ; питание органическим и неорганическим субстратом; факультативные анаэробы; обитают на поверхности
<b>Archaeoglobales-Methanomicrobia</b>	1,91	<b>Archaeoglobales:</b> оптимум $T=55-90^{\circ}\text{C}$ ; питание неорганическим субстратом; анаэробы; галофобы
		<b>Methanomicrobia:</b> оптимум $T=0-70^{\circ}\text{C}$ ; питание неорганическим субстратом; анаэробы; галофобы
<b>Halobacteriales-Methanomicrobia</b>	1,72	<b>Halobacteriales:</b> оптимум $T=31-50^{\circ}\text{C}$ ; питание органическим субстратом; аэробы; галофилы
		<b>Methanomicrobia:</b> оптимум $T=0-70^{\circ}\text{C}$ ; питание неорганическим субстратом; анаэробы; галофобы

Наибольшие значения  $Z_{score}$  соответствуют наибольшим изменениям индексов EEI2. На примере первого соседства Halobacteriales-Methanomicrobia из Таблицы 7: Halobacteriales –

оптимум  $T=31-50^{\circ}\text{C}$ , питание органическим субстратом, аэробы, галофилы; *Methanomicrobia* – оптимум  $T=0-70^{\circ}\text{C}$ , питание неорганическим субстратом, анаэробы, галофобы. Как можно видеть, имеются значительные различия в условиях среды обитания: различия в ширине температурного диапазона, питание органическим и неорганическим субстратом, аэробные и анаэробные условия, различное отношение к концентрации соли в окружающей среде.

Таким образом, из Таблицы 7 видно, что при изменении условий обитания организмов, наблюдаются наиболее значительные перестройки в первичной структуре их генов, что свидетельствует о приспособливании к новым условиям. В результате этого процесса изменяется профиль эффективности элонгации трансляции генов, что, вероятно, говорит об изменении профиля экспрессии генов. В новых условиях организм начинает более эффективно экспрессировать гены, которые в прежних условиях обитания экспрессировались на низком уровне и наоборот.

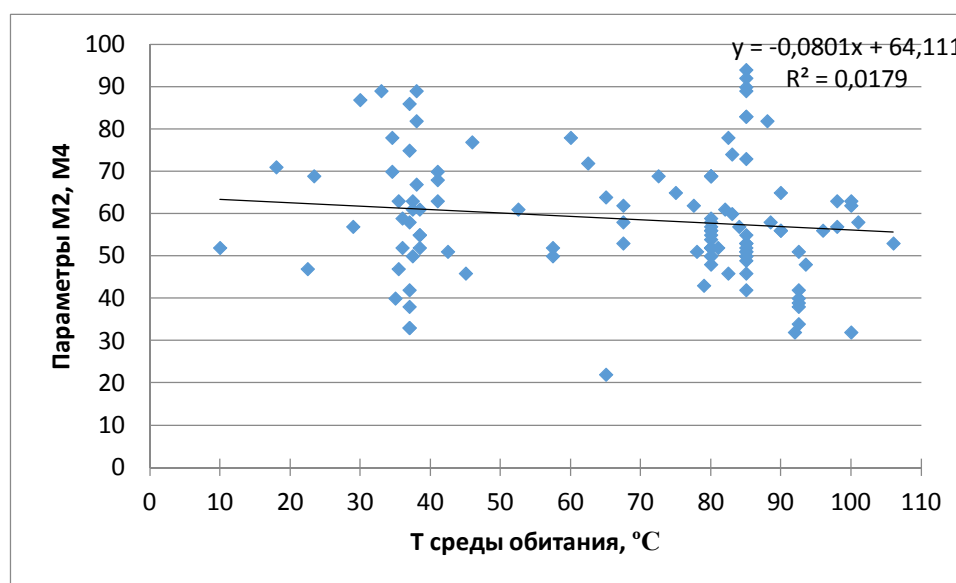
### **3.3.3. Исследование зависимости между влиянием потенциальных вторичных структур в мРНК на эффективность трансляции у архей и температурой их среды обитания**

Как было показано в пункте 3.1.2. главы «Результаты и обсуждение», в большинстве архей работают 2 и 4 типы индекса EEI. Другими словами, потенциальные вторичные структуры в мРНК являются одними из основных факторов, влияющих на эффективность элонгации трансляции генов. Однако стабильность вторичных структур зависит от температурных условий. Чем выше температура, тем менее стабильна структура. Поскольку археи являются пойкилотермными организмами (температура в клетке меняется в зависимости от температуры окружающей среды), низкая температура обитания потенциально может увеличить влияние вторичных структур на эффективность трансляции, а высокая – уменьшить. Для ответа на вопрос, связана ли температура среды обитания организма с силой влияния вторичных структур в мРНК на эффективность элонгации трансляции, был рассчитан коэффициент корреляции между данной температурой и параметром  $M$  у архей.

Данные по температурам мест обитания 135 архей были получены в основном из различных статей [Amo et al., 2002; Antunes et al., 2008; Arab et al., 2000; Birrien et al., 2011; Blöchl et al., 1997; Bonch-Osmolovskaya et al., 1990; Borges et al., 2010; Burns et al., 2007; Burns et al., 2010; Cadillo-Quiroz et al., 2014; Castillo et al., 2007; Chan et al., 2013; De Rosa et al., 1975; Dopson et al., 2004; Elkins et al., 2008; Erauso et al., 1993; Fan et al., 2004; Fiala et al., 1986; Fiala and Stetter 1986; González et al., 1998; Gruber et al., 2004; Gumerov et al., 2011; Gutiérrez et al., 2007; Hafenbradl et al., 1996; Huber et al., 1989; Huber et al., 1997; Huber et al., 1998; Huber et al., 2006; Itoh et al., 1999; Itoh et al., 2002; Itoh et al., 2003; Jahn et al., 2004; Jahn et al., 2007; Jolivet et al., 2003; Klenk et al., 1997; Kublanov et al., 2009; Liu et al., 2011; Mardanov et al., 2012; McGenity

et al., 1998; Miroshnichenko et al., 2001; Montalvo-Rodriguez et al., 1998; Nakagawa et al., 2004; Neuner et al., 1990; Niederberger et al., 2006; Oren et al., 1990; Oren et al., 2002; Perevalova et al., 2005; Perevalova et al., 2010; Preston et al., 1996; Prokofeva et al., 2009; Roh et al., 2007; Sako et al., 1996; Sako et al., 2001; Schleper et al., 1996; Schouten et al., 2008; Seob et al., 2006; Steinsbu et al., 2010; Suzuki et al., 2002; Torreblanca et al., 1986; Völkl et al., 1993; von Jan et al., 2010; Wainø et al., 2000; Yasuda et al., 1995; You et al., 2011; Zillig et al., 1991]. Часть данных была взята из базы данных по Methanogens (<http://metanogen.biotech.uni.wroc.pl/>), часть – с сайта <https://microbewiki.kenyon.edu/index.php/MicrobeWiki>. В качестве температуры среды обитания было взято среднее значение промежутка оптимальных температур роста организма. Полные данные по температуре среды обитания различных архей представлены в Таблице 7 в Приложении.

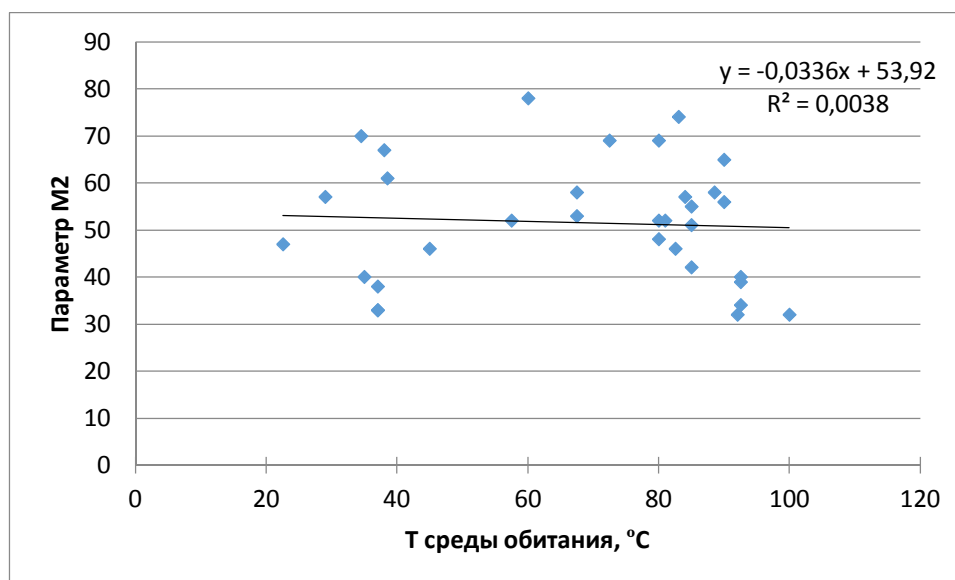
Результаты анализа представлены на Рисунках 51-53 в виде графиков зависимости параметра М от температуры среды обитания организма. На Рисунке 51 приведены данные для архей, у которых работает второй, либо четвертый тип ЕЕI. Значение коэффициента корреляции Спирмена –  $r = -0,15$  ( $P < 0,11$ ).



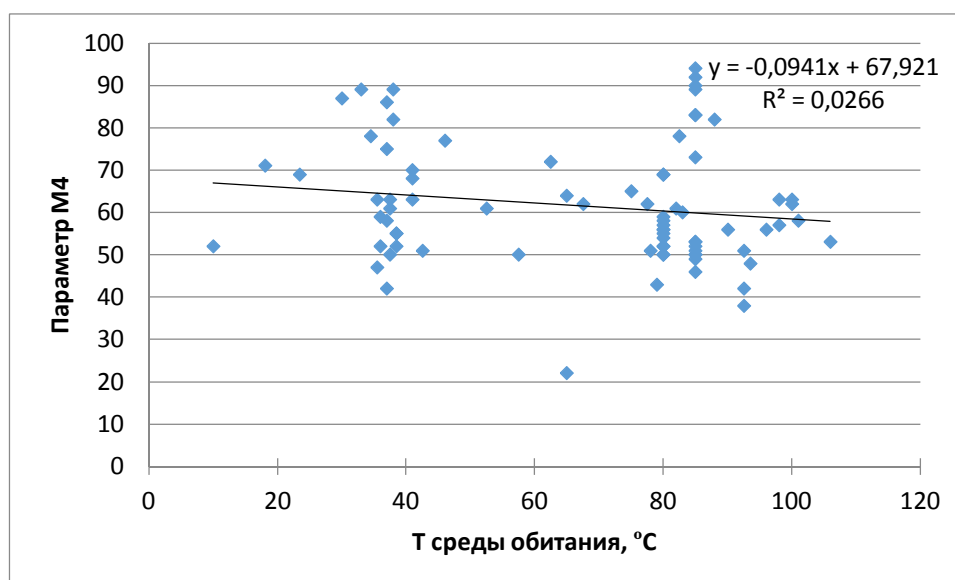
**Рис. 51.** Зависимость параметра М от температуры обитания архей (организмы, в которых работает 2 и 4 тип индекса ЕЕI) (Спирмен  $r = -0,15$ ;  $P < 0,11$ ).

На Рисунке 52 приведены данные для архей, у которых работает только второй тип ЕЕI. Значение коэффициента корреляции Спирмена –  $r = -0,19$  ( $P < 0,28$ ).

На Рисунке 53 приведены данные для архей, у которых работает только четвертый тип ЕЕI. Значение коэффициента корреляции Спирмена –  $r = -0,16$  ( $P < 0,17$ ).



**Рис. 52.** Зависимость параметра М от температуры обитания архей (организмы, в которых работает 2 тип индекса EEI) (Спирмен  $r = -0,19$ ;  $P < 0,28$ ).



**Рис. 53.** Зависимость параметра М от температуры обитания архей (организмы, в которых работает 4 тип индекса EEI) (Спирмен  $r = -0,16$ ;  $P < 0,17$ ).

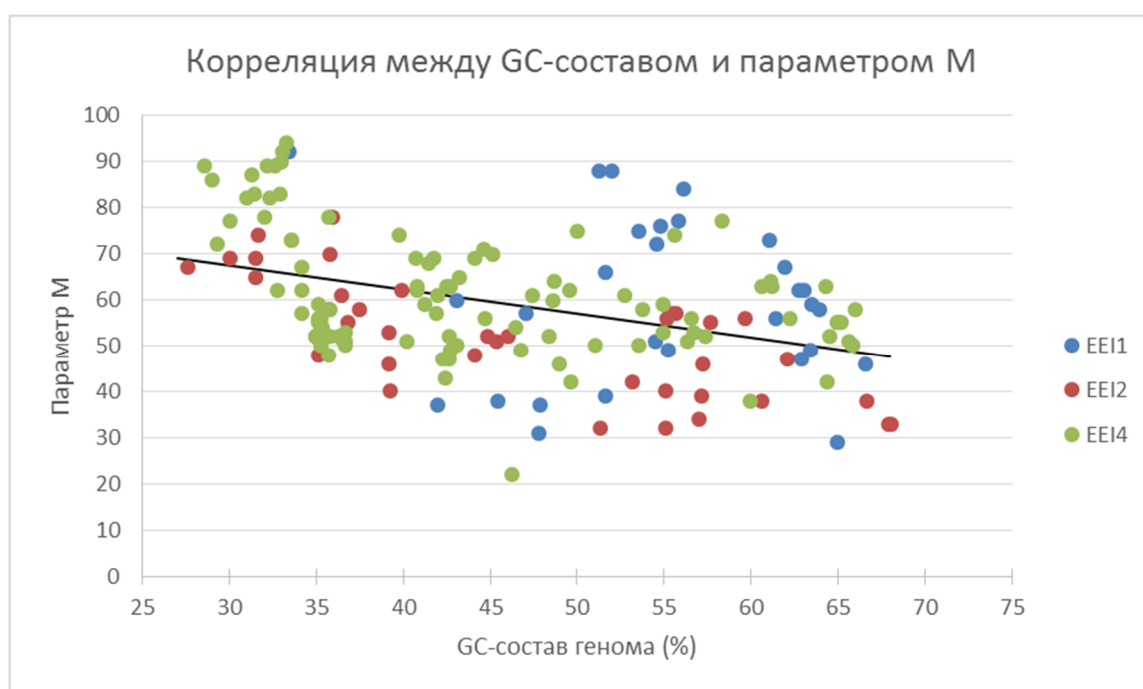
Также был проведен дисперсионный анализ. Археи были разделены на 3 группы по температуре среды обитания: мезофилы ( $T \leq 45^\circ\text{C}$ ), термофилы ( $45^\circ\text{C} < T \leq 60^\circ\text{C}$ ) и гипертермофилы ( $T > 60^\circ\text{C}$ ). Достоверных различий по параметру М среди групп обнаружено не было ( $P < 0,936$ ).

Таким образом, достоверной корреляции между температурой среды обитания архей и силой влияния потенциальных вторичных структур в мРНК на эффективность элонгации трансляции не обнаружено. Это может быть связано с тем, что археи, по-видимому, лишены энергозависимого механизма расплетания вторичных структур в мРНК, как было показано в разделе 3.1.2. главы «Результаты и обсуждение». Т.е. на эффективность трансляции их генов оказывает влияние количество потенциальных вторичных структур в мРНК и не оказывает их

энергетическая стабильность. Поскольку температура, в основном, влияет только на стабильность вторичных структур, это может быть причиной отсутствия корреляции. Однако этот вопрос требуют дополнительного исследования, в том числе экспериментального.

### 3.3.4. Связь между GC-составом и эволюционной оптимизацией первичной структуры генов архей для повышения эффективности элонгации трансляции

Для проверки зависимости эффективности элонгации трансляции от GC-состава генов у архей был построен график зависимости параметра М (наибольшее значение из М1-М5), показывающего эффективность определения ГРБ как высокоэкспрессируемых, от GC-состава генома данных организмов (Рисунок 54).



**Рис. 54.** Зависимость параметра М от GC-состава генома для 135 видов архей (Спирмен  $r = -0,40$ ;  $P < 8,65 \times 10^{-8}$ ). Синим цветом выделены организмы, в которых работает EEI1, красным – EEI2, зеленым – EEI4. Черная линия – линия регрессии.

Имеет место достоверная отрицательная корреляция между GC-составом генома и степенью эволюционной оптимизации первичной структуры его генов для повышения эффективности элонгации трансляции (параметр М). Коэффициент корреляции Спирмена  $r = -0,40$  ( $P < 8,65 \times 10^{-8}$ ). Данный результат хорошо согласуется с литературными данными о том, что GC-состав генома является одним из важных факторов, определяющих эффективность трансляции (раздел 1.2.1.2. главы «Обзор литературы»).

### 3.4. Исследование взаимосвязи между эффективностью элонгации трансляции генов дрожжей и плотностью их нуклеосомной упаковки в 5'-фланкирующем районе

Как известно, ДНК в клетках эукариот упакована при помощи специальных белков гистонов, образующих комплексы, называемые нуклеосомами. Для прохождения транскрипции ДНК с последовательностью гена должна быть освобождена от нуклеосом. Поскольку высокоэкспрессирующиеся гены должны обладать высокой эффективностью не только трансляции, но и транскрипции, плотность их нуклеосомной упаковки должна быть несколько ниже, чем у остальных генов.

Для оценки согласованности процессов на разных уровнях экспрессии генов было проведено исследование корреляции между плотностью нуклеосомной упаковки в 5'-фланкирующих районах генов дрожжей видов *S. cerevisiae* и *S. pombe* и значениями индекса эффективности элонгации ЕЕI соответствующих генов.

Проверяемая гипотеза заключалась в следующем: для эффективной экспрессии генов необходимы согласованно оптимизированные процессы транскрипции и трансляции, в частности – инициации транскрипции и элонгации трансляции. Такая корреляция была найдена между потенциалом формирования нуклеосом (ПФН) в 5'-фланкирующем районе генов дрожжей видов *S. cerevisiae* и *S. pombe* и их значениями ЕЕI.

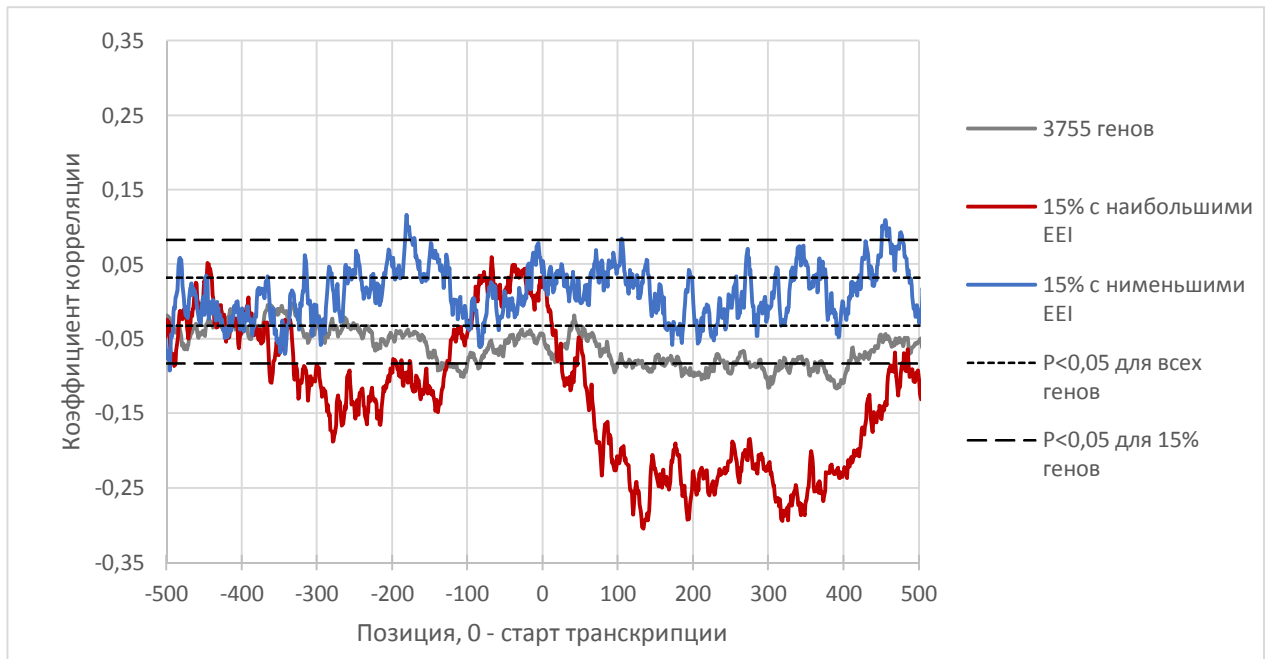
Для 5'-фланкирующего района каждого гена был рассчитан потенциал формирования нуклеосом – функция, которая характеризует вероятность расположения нуклеосомы в заданном сайте последовательности. Значения этой функции вычисляются на основе частот динуклеотидов. Для расчета ПФН использовалась программа RECON (<http://www.mgs.bionet.nsc.ru/mgs/programs/recon>) [Levitsky, 2004]. Эта программа рассчитывает нуклеосомный потенциал в окне шириной  $W = 140$  нт по всей последовательности длины  $L$ . Значение потенциала приписывается позиции центра окна. При анализе одной последовательности получается профиль ПФН длины  $L - W + 1$ . Значения ПФН меньше 1, а значение 1 соответствует наибольшей вероятности наличия сайта формирования нуклеосомы.

#### 3.4.1. Корреляция между потенциалом формирования нуклеосом и ЕЕI у *S. pombe*

Как показано выше, ЕЕI применим для оценки эффективности элонгации трансляции генов *S. cerevisiae* и *S. pombe* (Таблица 3, раздел 3.1.3. главы «Результаты и обсуждение»). Для *S. cerevisiae* коэффициент корреляции между ЕЕI и экспериментально определенной эффективностью экспрессии составляет  $r = 0,79$  ( $P < 10^{-5}$ ) [Владимиров и др., 2007].

Для 3755 генов *S. pombe* (75% от общего количества) с известными позициями стартов транскрипции [Lantermann et al., 2010] были рассчитаны коэффициенты корреляции Спирмена

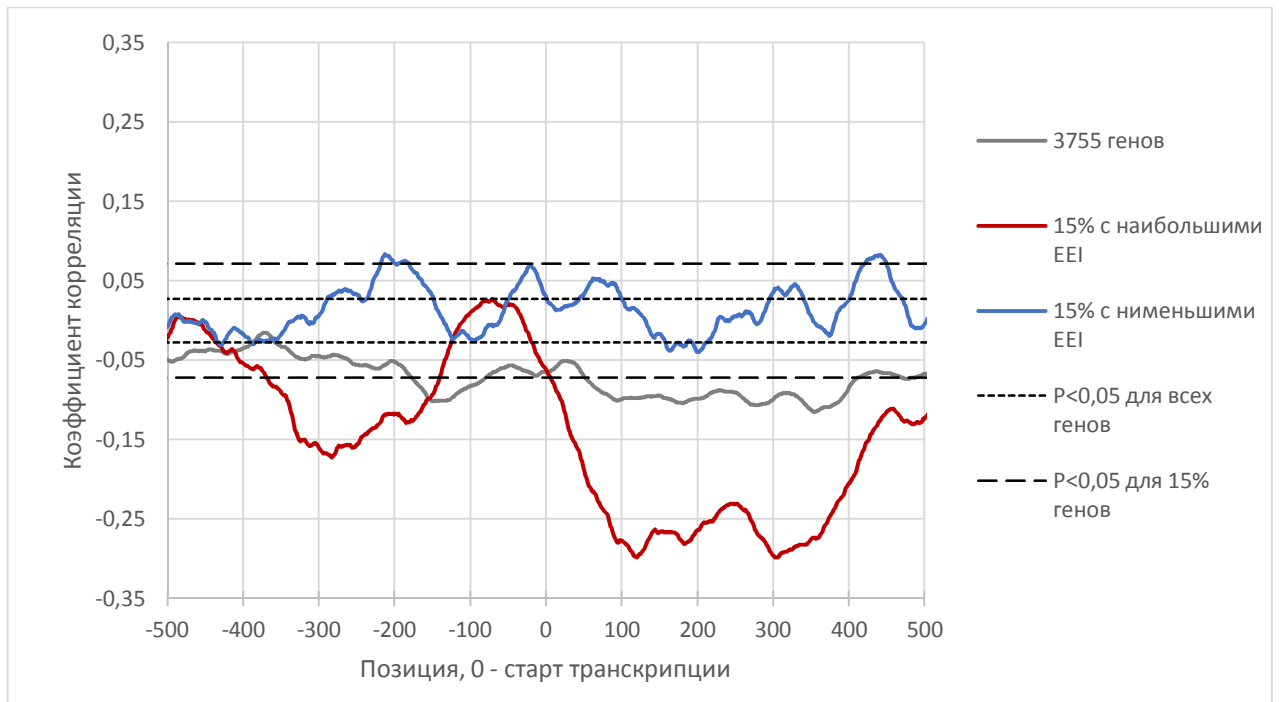
между двумя векторами: вектором значений ПФН для определенной позиции фазированных относительно старта транскрипции всех последовательностей генов и вектором значений EEI этих последовательностей. Профили коэффициентов корреляции между ПФН и EEI были получены для участков  $[-500; +500]$  относительно старта транскрипции генов. Результаты представлены на Рисунке 55.



**Рис. 55.** Корреляция между EEI и ПФН для *S. pombe*: для 3755 последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует началу транскрипции. Горизонтальные линии – зоны достоверности ( $P < 0,05$ ) по критерию Фишера.

Как можно видеть, имеет место достоверная отрицательная корреляция между EEI и ПФН для высокоэкспрессирующихся последовательностей в интервале  $[-330; -130]$ . Т.е., чем слабее нуклеосомная упаковка в 5'-области, тем выше эффективность элонгации трансляции мРНК. В районе ниже старта транскрипции  $[0; +500]$  наблюдается достоверная отрицательная корреляция для высокоэкспрессирующихся последовательностей. Видимо, это связано с особенностями нуклеотидного и динуклеотидного состава кодирующих частей генов (см. далее).

Также для *S. pombe* были рассчитаны коэффициенты корреляции Спирмена между двумя векторами: вектором средних значений ПФН для нескольких позиций в окне шириной 50 нуклеотидов фазированных относительно старта транскрипции всех последовательностей генов и вектором значений EEI этих последовательностей. Результаты представлены на Рисунке 56. Как можно видеть, интервалы достоверных значений корреляции аналогичны интервалам на Рисунке 55. Это еще раз подтверждает связь между плотностью нуклеосомной упаковки генов в данных районах и эффективностью их элонгации трансляции у *S. pombe*.



**Рис. 56.** Корреляция между EEI и средним значением ПФН для окна шириной 50 нуклеотидов для *S. pombe*: для 3755 последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует началу транскрипции. Горизонтальные линии – зоны достоверности ( $P < 0,05$ ) по критерию Фишера.

### 3.4.2. Корреляция между потенциалом формирования нуклеосом и EEI у *S. cerevisiae*

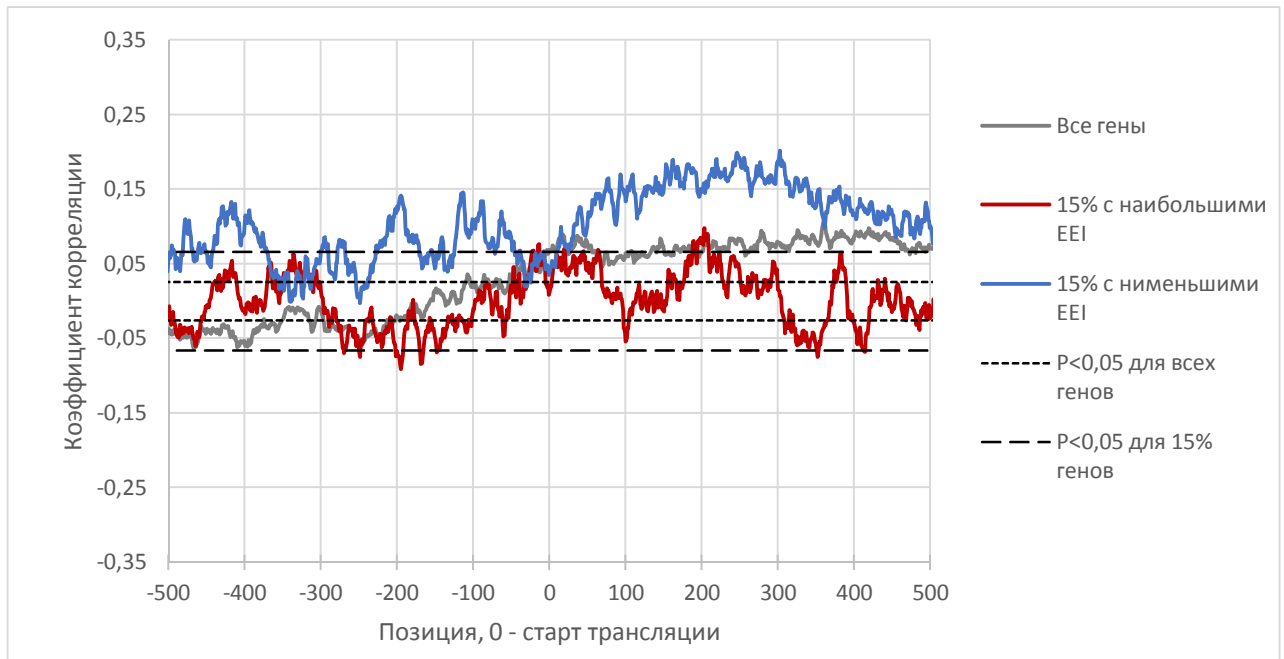
Для 5829 генов *S. cerevisiae* были рассчитаны коэффициенты корреляции Спирмена между двумя векторами: вектором значений ПФН для определенной позиции фазированных относительно старта трансляции (AUG) всех последовательностей генов и вектором значений EEI этих последовательностей. Профили коэффициентов корреляции между ПФН и EEI были получены для участков  $[-500; +500]$  относительно старта трансляции генов. Результаты представлены на Рисунке 57.

Как видно из Рисунка 57, есть достоверная положительная корреляция между EEI и ПФН для низкоэкспрессирующихся последовательностей в интервалах  $[-120; -40]$ ,  $[-230; -170]$ ,  $[-500; -370]$ . Т.е., чем крепче нуклеосомная упаковка в 5'-области, тем выше эффективность элонгации трансляции мРНК. Как будет показано далее, высоко достоверная положительная корреляция между EEI и ПФН для низкоэкспрессирующихся генов в начале кодирующей части  $[0; +500]$ , видимо, связана с нуклеотидным и динуклеотидным составом этих последовательностей.

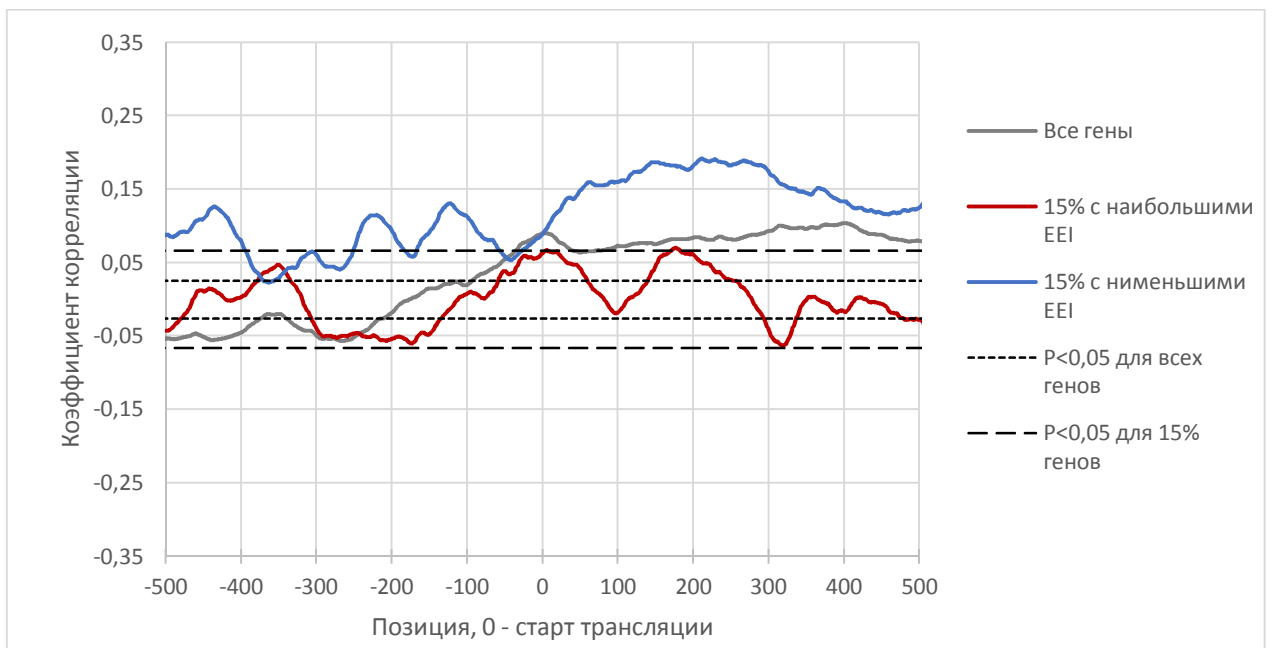
Также для *S. cerevisiae* были рассчитаны коэффициенты корреляции Спирмена между двумя векторами: вектором средних значений ПФН для нескольких позиций в окне шириной 50 нуклеотидов фазированных относительно старта трансляции (AUG) всех последовательностей



генов и вектором значений EEI этих последовательностей. Результаты представлены на Рисунке 58.



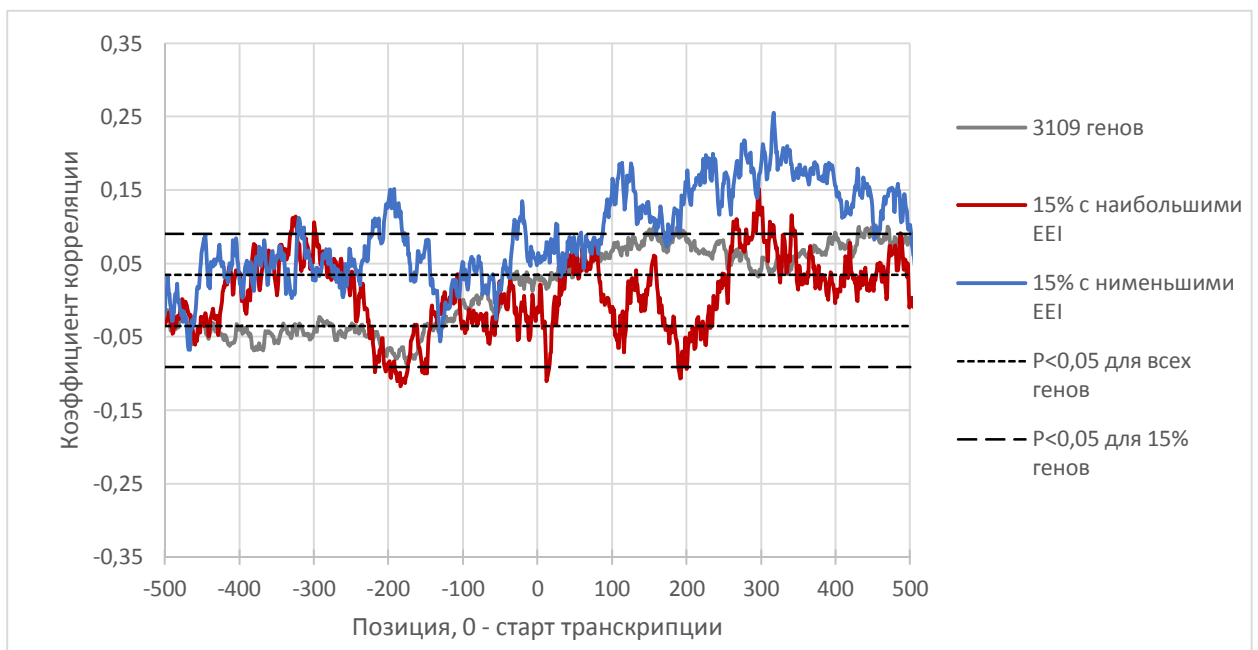
**Рис. 57.** Корреляция между EEI и ПФН для *S. cerevisiae*: для всех последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует положению нуклеотида А в старте-кодоне трансляции АТГ. Горизонтальные линии – зоны достоверности ( $P < 0,05$ ) по критерию Фишера.



**Рис. 58.** Корреляция между EEI и средним значением ПФН для окна шириной 25 нуклеотидов для *S. cerevisiae*: для всех последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует положению нуклеотида А в старте-кодоне трансляции АТГ. Горизонтальные линии – зоны достоверности ( $P < 0,05$ ) по критерию Фишера.

Как видно из Рисунка 58, интервалы достоверных значений корреляции аналогичны интервалам на Рисунке 57. Это еще раз подтверждает связь между плотностью нуклеосомной упаковки генов в данных районах и эффективностью их элонгации трансляции у *S. cerevisiae*.

Также для 3109 генов *S. cerevisiae* (53% от общего количества) с известными позициями стартов транскрипции [Xia et al., 2011] были рассчитаны коэффициенты корреляции Спирмена между двумя векторами: вектором значений ПФН для определенной позиции фазированных относительно старта транскрипции всех последовательностей генов и вектором значений EEI этих последовательностей. Профили коэффициентов корреляции между ПФН и EEI были получены для участков [-500; +500] относительно старта транскрипции генов. Результаты представлены на Рисунке 59.

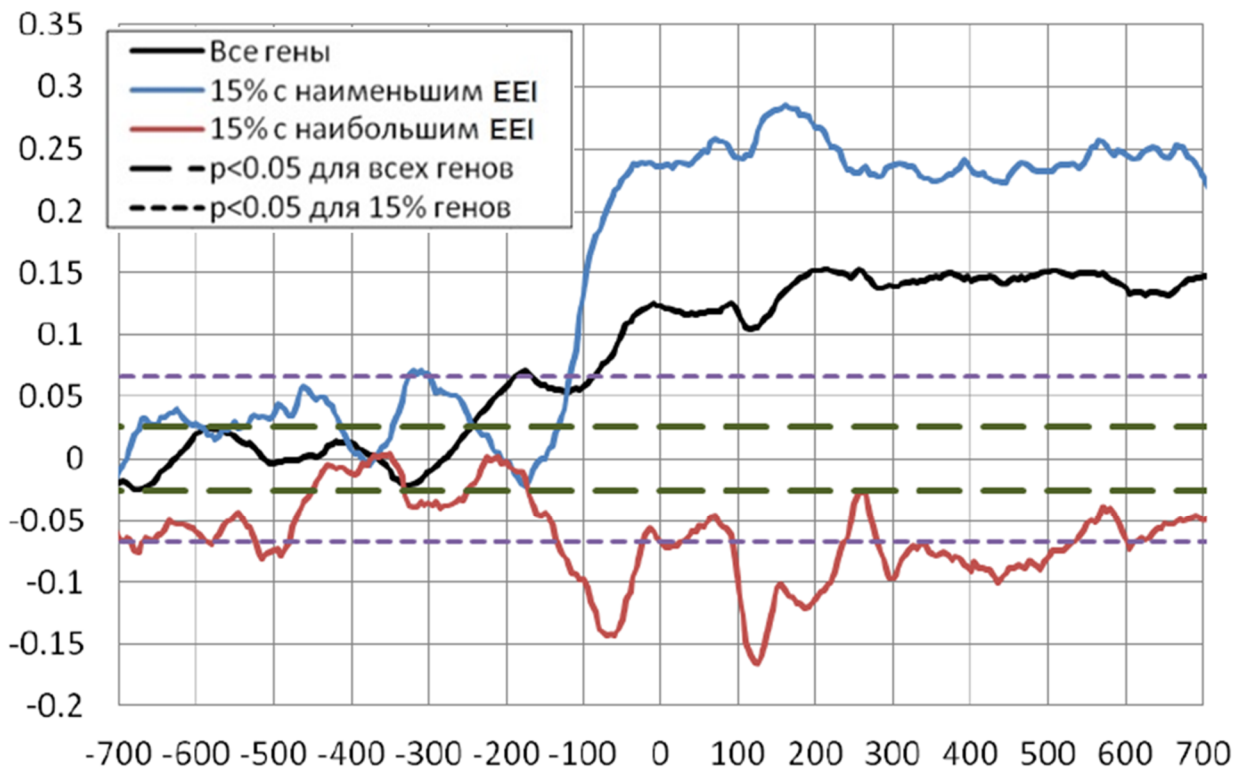


**Рис. 59.** Корреляция между EEI и ПФН для *S. cerevisiae*: для 3109 последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует началу транскрипции. Горизонтальные линии – зоны достоверности ( $P < 0,05$ ) по критерию Фишера.

Как можно видеть из Рисунка 59, есть достоверная положительная корреляция между EEI и ПФН для низкоэкспрессирующихся последовательностей в интервале [-220; -180]. Также есть достоверная положительная корреляция между EEI и ПФН для низкоэкспрессирующихся генов в районе ниже старта транскрипции [0; +500]. Результаты частично совпадают с данными, полученными при выравнивании последовательностей по старт-кодону трансляции. Неполное совпадение, возможно, связано с недостаточно большой выборкой генов с известными стартами транскрипции (53%).

### 3.4.3. Корреляция между EEI и экспериментальными данными по нуклеосомной упаковке у *S. cerevisiae*

Для подтверждения теоретических результатов были рассчитаны коэффициенты корреляции между EEI генов *S. cerevisiae* и экспериментальными данными по расположению нуклеосом в геномной ДНК (Рисунок 60) [Kaplan et al., 2009]. Короткие фрагменты ДНК, связанные с нуклеосомой, были наложены на геном. Также были отфильтрованы полностью идентичные фрагменты и фрагменты, не имеющие однозначной локализации в геноме. Удлиненные до 146 нуклеотидов фрагменты, что соответствует размеру нуклеосомы, формируют наложенный друг на друга профиль в хромосомных координатах, который и является плотностью, или картой расположения нуклеосом.



**Рис. 60.** Коэффициенты корреляции (ось ординат) между EEI и экспериментальными данными по плотности нуклеосомной упаковки, полученной прямым секвенированием нуклеосомной ДНК для *S. cerevisiae*: для всех последовательностей, для 15% высокоэкспрессирующихся по EEI и 15% низкоэкспрессирующихся по EEI. Позиция 0 соответствует положению нуклеотида А в старт-кодоне трансляции АТГ. Горизонтальные линии – зоны достоверности ( $P < 0,05$ ) по критерию Фишера.

Таким образом, каждой позиции на хромосоме соответствует число фрагментов секвенированной нуклеосомной ДНК, перекрывающих данную позицию. Это число, соответствующее экспериментальной плотности расположения нуклеосом, может быть сравнено с любым другим, характеризующим данную позицию, например, с полученным на основе анализа локального контекста, либо экспериментально.

Как видно из Рисунка 60, для низкоэкспрессирующихся последовательностей в интервале  $[-110; 0]$  имеет место достоверная положительная корреляция ( $r \leq 0,25$ ). Этот результат совпадает с теоретическими данными (Рисунки 57-58), что подтверждает их достоверность. Для высокоэкспрессирующихся последовательностей в этом же интервале корреляция достоверна и отрицательна ( $r \geq -0,15$ ).

Результаты подтверждают проверяемую гипотезу, хотя по локализации областей значимых корреляций не вполне совпадают с найденными исходя из теоретических оценок. Отличие нуклеосомной локализации, определенной по экспериментальным данным *in vivo*, от теоретически предсказанной в промоторах, особенно в районе  $[-200; 0]$ , было также показано в статьях [Kaplan, 2009; Goh, 2010].

### **3.5. Исследование возможных причин различия видов корреляций между ПФН и ЕЕІ у *S. cerevisiae* и *S. pombe***

#### **3.5.1. Исследование профилей нуклеосомного потенциала у *S. cerevisiae* и *S. pombe***

Как видно из Рисунков 55-59, имеет место высокое и достоверное значение коэффициентов корреляции между потенциалом формирования нуклеосом (ПФН) и индексом эффективности элонгации трансляции (ЕЕІ) в кодирующей части генов дрожжей в районе  $[0; +500]$ . Также заметно явное различие в значениях коэффициентов корреляций для высоко- и низкоэкспрессирующихся генов *S. cerevisiae* и *S. pombe*. Для выяснения причин данных различий между организмами был проведен подробный анализ кодирующих частей их генов и геномов в целом.

Для начала были построены графики самих ПФН, усредненные по всем генам исследуемых организмов и отдельно по 15% генов с наибольшими и наименьшими значениями ЕЕІ (Рисунки 61-62). Если сравнить эти графики с представленными на Рисунках 55-58, можно заметить, что наибольшее значение коэффициента корреляции наблюдается при наименьших значениях ПФН. Коэффициент корреляции Пирсона/Спирмена между средними значениями ПФН и значениями коэффициента корреляции между ПФН и ЕЕІ для интервала  $[-600; +600]$  равен  $r = -0,68/-0,67$  ( $P < 1,04 \times 10^{-173}/1,04 \times 10^{-173}$ ) для *S. cerevisiae* и  $r = 0,72/0,55$  ( $P < 1,04 \times 10^{-173}/1,04 \times 10^{-173}$ ) для *S. pombe*. Т.е. корреляция здесь действительно присутствует и достаточно высока. Таким образом, районы с наименьшим потенциалом формирования нуклеосом, т.е. с наименьшей вероятностью их образования, оказывают наибольшее влияние на определение скорости элонгации трансляции.

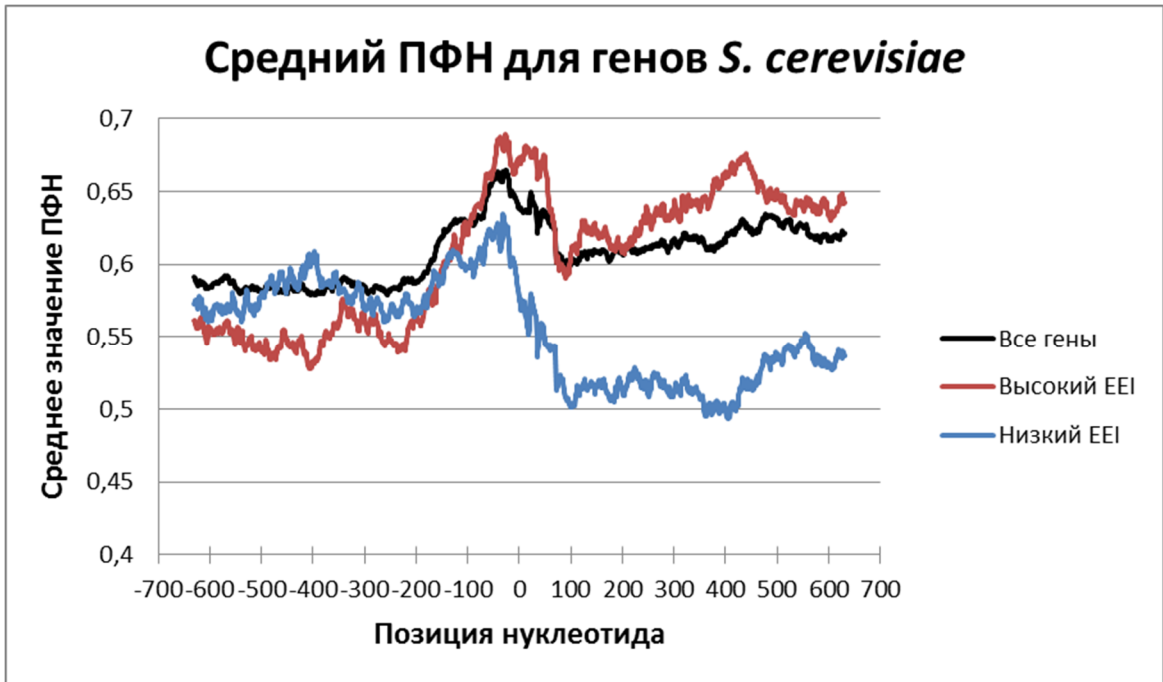


Рис. 61. Средний ПФН по всем генам, 15% с наибольшими значениями EEI, 15% с наименьшими значениями EEI в *S. cerevisiae* (0 – старт-кодон).

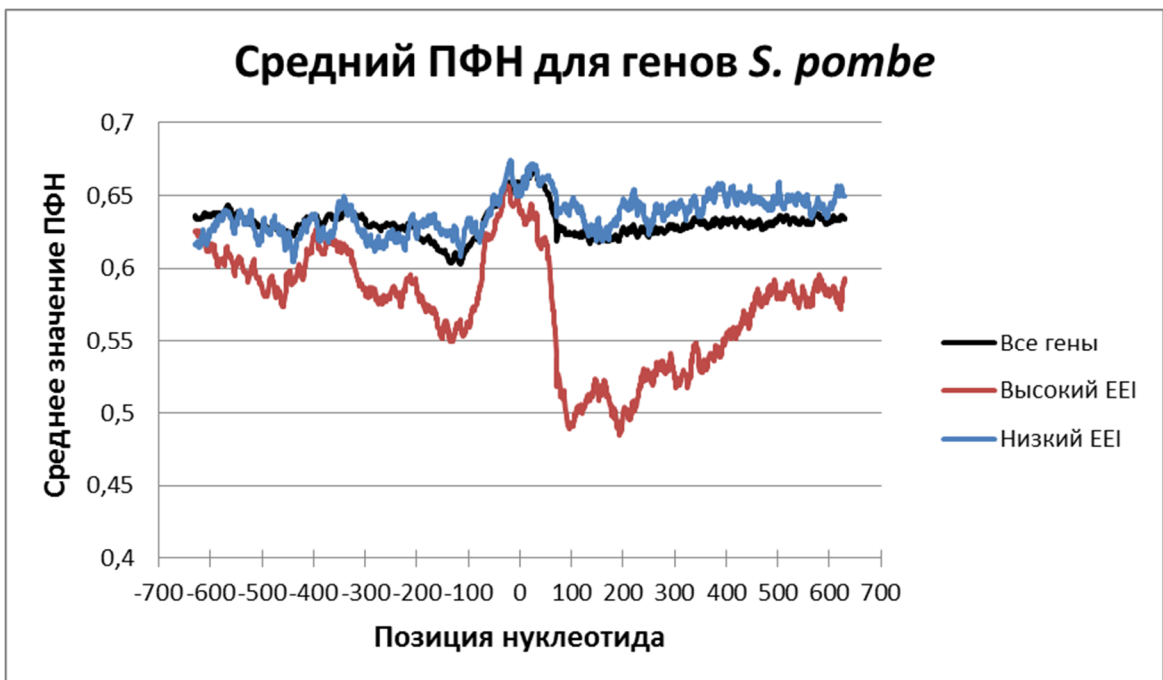


Рис. 62. Средний ПФН по всем генам, 15% с наибольшими значениями EEI, 15% с наименьшими значениями EEI в *S. pombe* (0 – старт-кодон).

### 3.5.2. Распределение генов по длине у *S. cerevisiae* и *S. pombe*

Для определения длины кодирующей части большинства генов на интервале [0; +600] были построены гистограммы распределения длин генов *S. cerevisiae* и *S. pombe*. В результате показано, что для  $\geq 90\%$  генов интервал [0; +400] является кодирующей частью (Рисунки 63-64).

Таким образом, стоит рассматривать корреляцию между ПФН и ЕЕІ только до позиции +400 относительно старта трансляции.

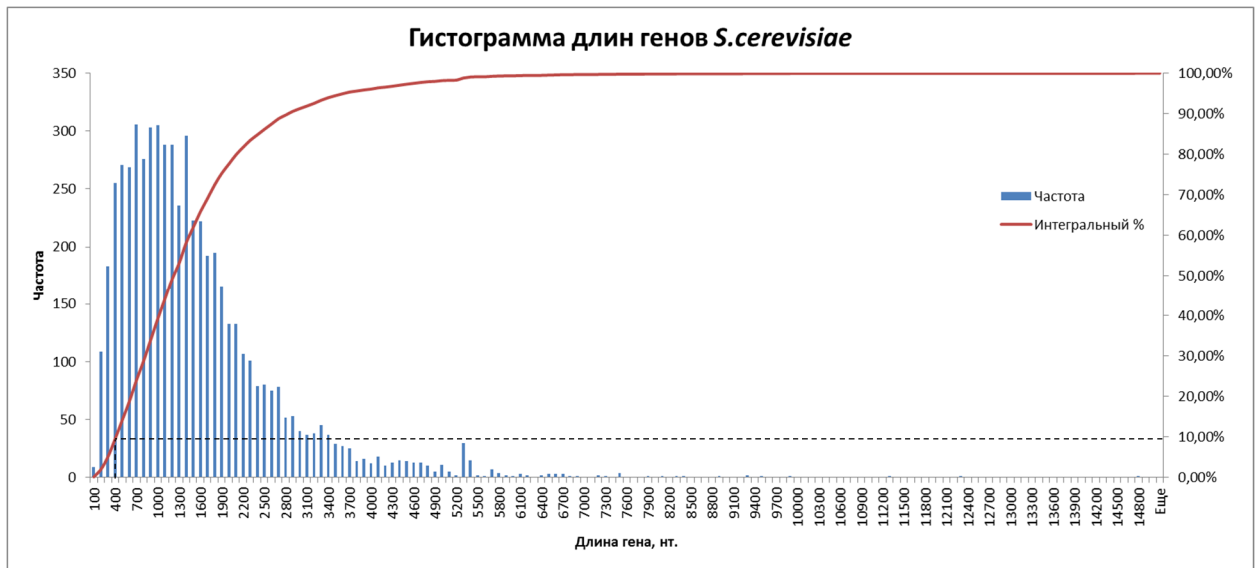


Рис. 63. Гистограмма длин генов *S. cerevisiae*.

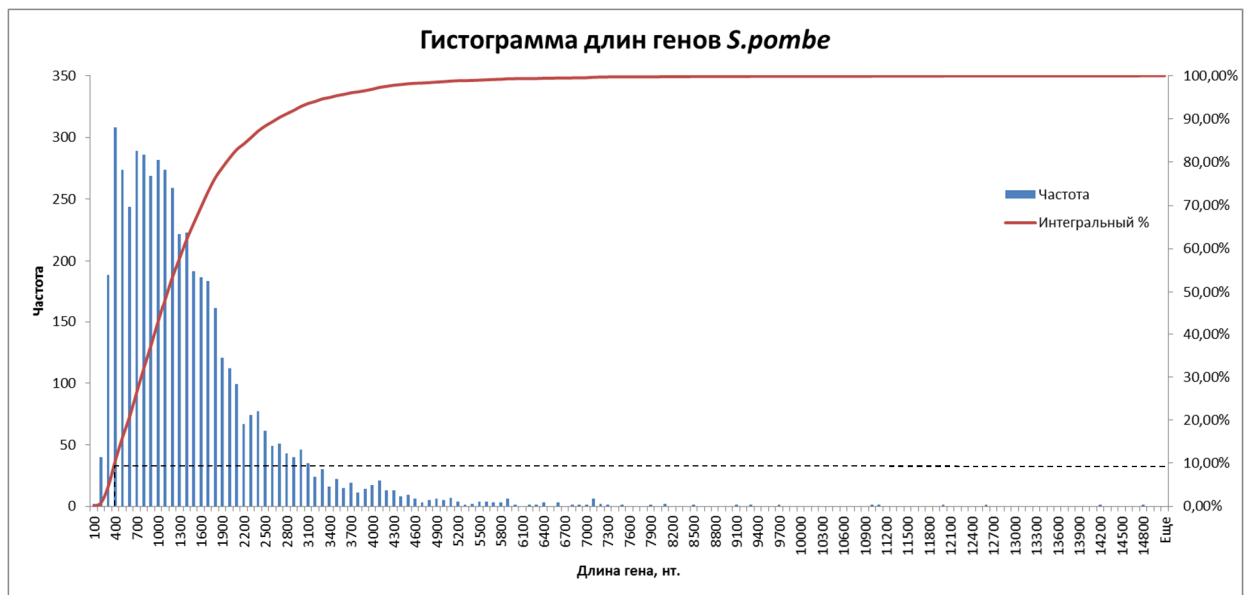
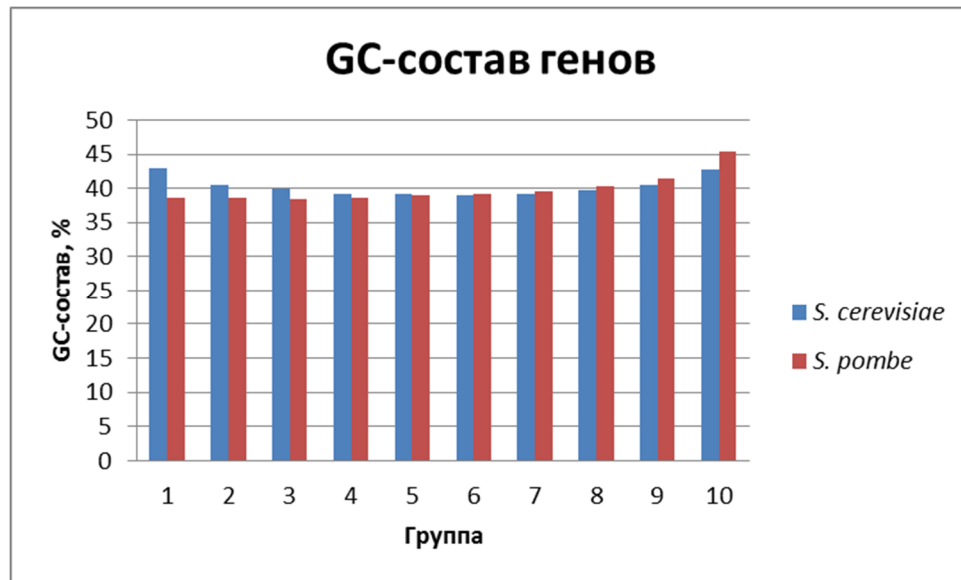


Рис. 64. Гистограмма длин генов *S. pombe*.

### 3.5.3. Исследование связи между GC-составом генов и индексом ЕЕІ у *S. cerevisiae* и *S. pombe*

На Рисунке 65 приведены значения GC-состава генов *S. cerevisiae* и *S. pombe*. Все гены данных организмов были упорядочены по возрастанию индекса ЕЕІ и разделены на 10 групп по 10% в каждой. Для каждой группы генов был подсчитан их средний GC-состав.



**Рис. 65.** GC-состав генов *S. cerevisiae* и *S. pombe*, разделенных на 10 групп по значению EEI. В 1 группе – гены с наименьшими значениями EEI, в 10 – с наибольшими.

Геномы исследуемых организмов мало отличаются по GC-составу: 38,30% у *S. cerevisiae* и 36,06% у *S. pombe*. Как видно из Рисунка 65, у *S. cerevisiae* группы высоко- и низкоэкспрессирующихся генов имеют примерно одинаковый GC-состав (42,71% и 42,94% соответственно). Из чего можно сделать вывод, что GC-состав гена не оказывает значительного влияния на эффективность элонгации трансляции. У *S. pombe* в свою очередь наблюдается обратная картина – группа высокоэкспрессирующихся генов имеет наибольшее значение GC-состава (45,39%), а группа низкоэкспрессирующихся – одно из наименьших (38,60%). В данном организме гены с большим значением GC-состава имеют более высокую скорость элонгации трансляции.

### 3.5.4. Сравнение кодонных составов генов *S. cerevisiae* и *S. pombe*

Поскольку в дрожжах примерно с одинаковой вероятностью могут работать как первый, так и четвертый тип индекса EEI, было проведено сравнение кодонных составов генов данных организмов, а также составов пулов тРНК по количеству тРНК на каждый кодон. Количество тРНК было оценено на основе количества кодирующих их генов, так как известно, что данные величины высоко и достоверно коррелируют.

Были исследованы: а) кодонный состав всех генов данных организмов, б) кодонный состав 15% высокоэкспрессирующихся генов, в) частоты генов тРНК, соответствующих определенным кодонам. Значения коэффициентов корреляции Спирмена между соответствующими величинами представлены в Таблице 8 (достоверность всех коэффициентов >99,9%).

Количество генов, кодирующих тРНК в геномах *S. cerevisiae* и *S. pombe*, равно 275 и 170 соответственно. То есть количество значительно различается. Из Таблицы 8 видно, что по распределению антикодонов тРНК *S. cerevisiae* и *S. pombe* коррелируют на величину  $r = 0,80$  ( $P < 7,23 \times 10^{-9}$ ). Таким образом, можно заключить, что пулы тРНК у данных организмов имеют некоторые различия, существенные для определения предпочтительных кодонов. Как и следовало ожидать, кодонные составы генов каждого из организмов лучше коррелируют с антикодонным составом своих тРНК, чем тРНК другого организма. Высокоэкспрессирующиеся гены по кодонному составу коррелируют на  $r = 0,79$  ( $P < 2,20 \times 10^{-16}$ ), а все гены организмов коррелируют на  $r = 0,91$  ( $P < 2,20 \times 10^{-16}$ ).

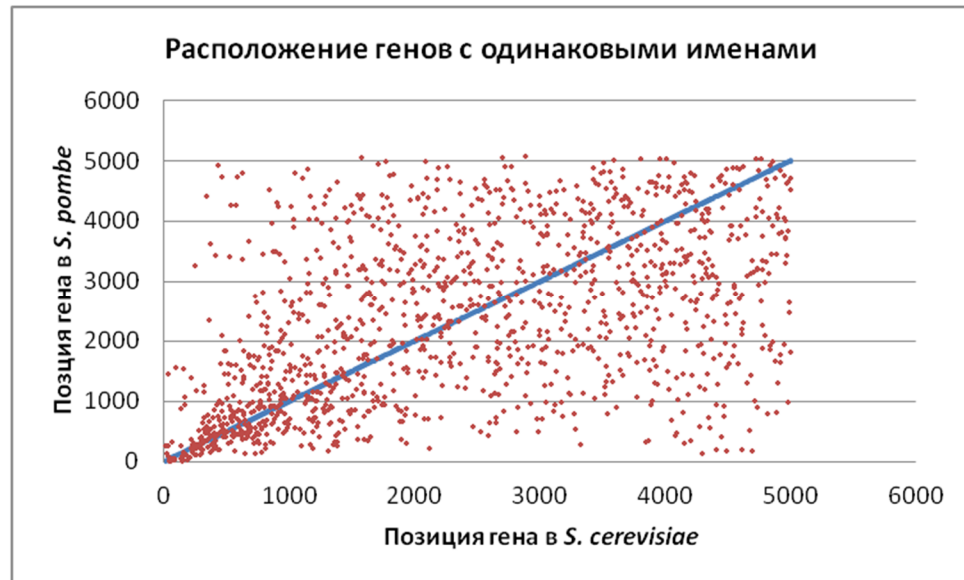
**Таблица 8.** Результаты анализа кодонных составов генов и антикодонных составов пулов тРНК *S. cerevisiae* и *S. pombe* (красный фон – корреляция между аналогичными группами генов разных организмов, зеленый – корреляция между разными группами генов одного организма).

		Все гены		Высоко-экспрессирующиеся		тРНК	
		<i>S.cerevisiae</i>	<i>S.pombe</i>	<i>S.cerevisiae</i>	<i>S.pombe</i>	<i>S.cerevisiae</i>	<i>S.pombe</i>
Все гены	<i>S.cerevisiae</i>	1,00	0,91	0,90	0,76	0,77	0,59
	<i>S.pombe</i>		1,00	0,79	0,90	0,68	0,71
Высоко-экспрессирующиеся	<i>S.cerevisiae</i>			1,00	0,79	0,83	0,63
	<i>S.pombe</i>				1,00	0,72	0,82
тРНК	<i>S.cerevisiae</i>					1,00	0,80
	<i>S.pombe</i>						1,00

### 3.5.5. Сравнение эффективностей элонгации трансляции генов с одинаковыми идентификаторами у *S. cerevisiae* и *S. pombe* при помощи программы EloE

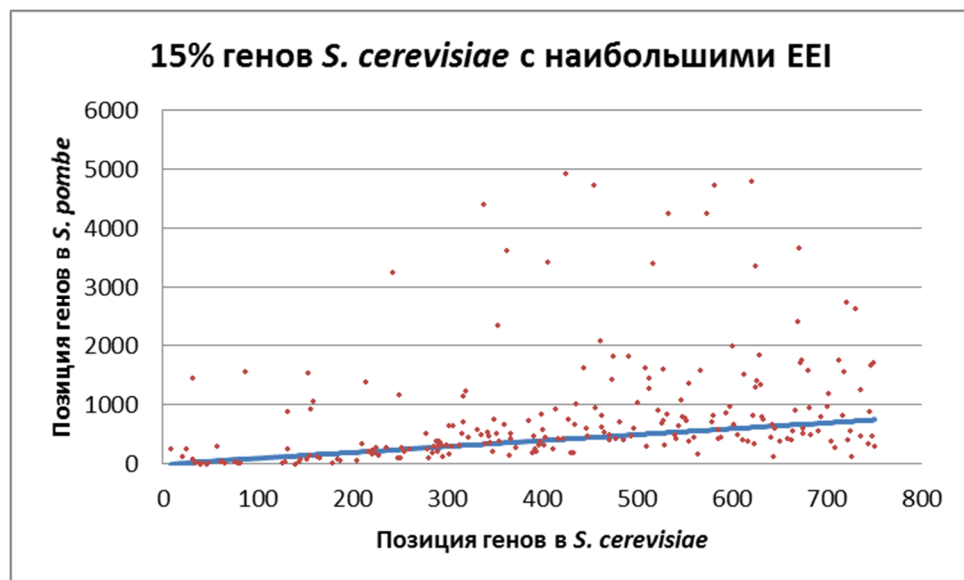
Поскольку пулы тРНК у *S. cerevisiae* и *S. pombe* имеют небольшие различия, а кодонные составы генов схожи, можно предположить, что ортологичные гены будут иметь разный уровень эффективности элонгации трансляции. На Рисунке 66 показано расположение генов (красные точки) с одинаковыми идентификаторами в отсортированных по уменьшению EEI1 списках генов *S. cerevisiae* и *S. pombe*.



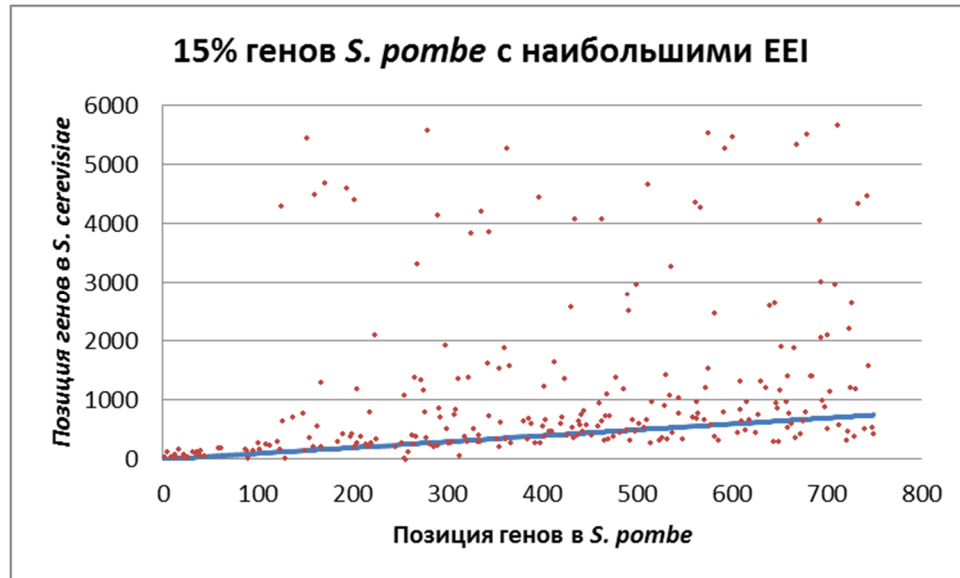


**Рис. 66.** Расположение генов (красные точки) с одинаковыми именами в отсортированных по уменьшению EEI списках генов *S. cerevisiae* и *S. pombe*. Синяя линия – линия равенства позиций генов в обоих списках.

На Рисунках 67-68 показаны 15% высокоэкспрессирующихся генов. Как видно из графиков, не все высокоэкспрессирующиеся гены из одного организма также являются высокоэкспрессирующимися в другом (красные точки, значительно удаленные от синей линии). Есть гены с одинаковыми идентификаторами, но разными уровнями эффективности элонгации трансляции. Возможно, это является следствием различий в пулах тРНК. Также это может быть связано с различиями по востребованности функций этих генов в каждом организме (см. раздел 1.2.1.3. главы «Обзор литературы»).



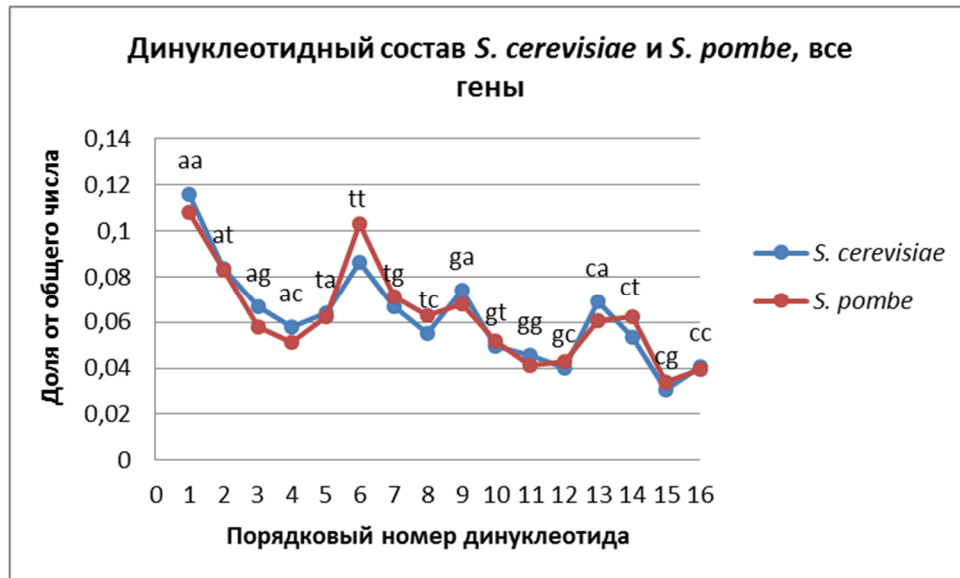
**Рис. 67.** Расположение 15% высокоэкспрессируемых генов из *S. cerevisiae* и их аналогов с одинаковыми именами в отсортированном по уменьшению EEI списке генов *S. pombe*. Синяя линия – линия равенства позиций генов в обоих списках.



**Рис. 68.** Расположение 15% высокоэкспрессируемых генов из *S. pombe* и их аналогов с одинаковыми именами в отсортированном по уменьшению EEI списке генов *S. cerevisiae*. Синяя линия – линия равенства позиций генов в обоих списках.

### 3.5.6. Сравнение динуклеотидных составов генов *S. cerevisiae* и *S. pombe*

Динуклеотидный состав генов может различаться у исследуемых организмов и тем самым влиять на вид корреляции между ПФН и EEI, т.к. расчет ПФН основан на учете динуклеотидного состава. Поэтому данный параметр также был проанализирован (Рисунки 69-71).



**Рис. 69.** Динуклеотидный состав всех генов *S. cerevisiae* и *S. pombe*.

Динуклеотидные составы всех генов исследуемых организмов имеют небольшие различия: динуклеотиды «ТТ» и «СТ» встречается в генах *S. cerevisiae* достоверно реже ( $P < 2,20 \times 10^{-16}$ ), чем у *S. pombe*, а «АА» и «СА» – чаще (Рисунок 69). Высокоэкспрессирующиеся

гены исследуемых организмов имеют наибольшие различия по динуклеотидному составу (Рисунок 70).

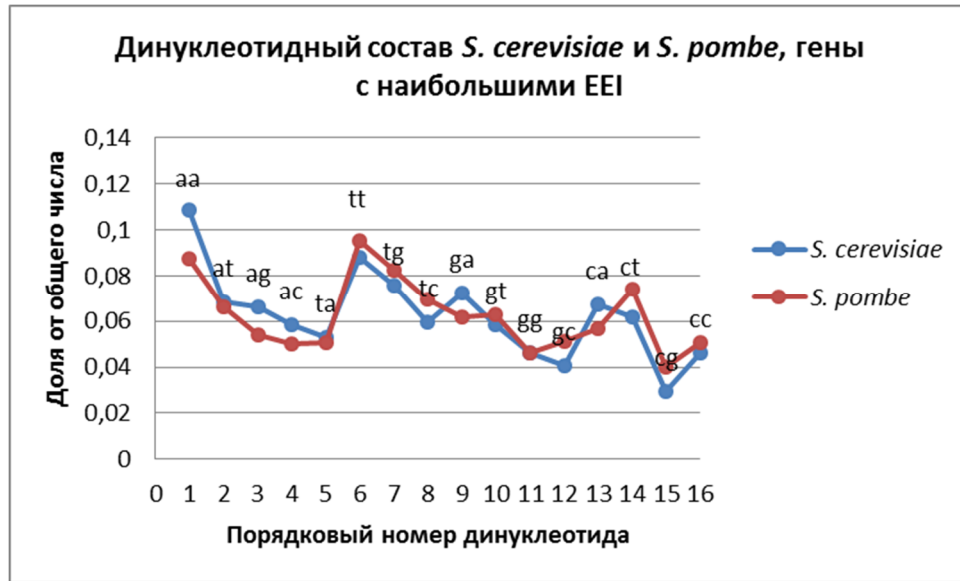


Рис. 70. Динуклеотидный состав 15% генов *S. cerevisiae* и *S. pombe* с наибольшими значениями EEI.

Результаты анализа в виде коэффициентов корреляции между составами динуклеотидов разных наборов генов представлены в Таблице 9 (достоверность большинства коэффициентов >99,9%, кроме «Высокоэкспрессирующиеся *S. pombe*» – «Низкоэкспрессирующиеся *S. cerevisiae*» >99%).

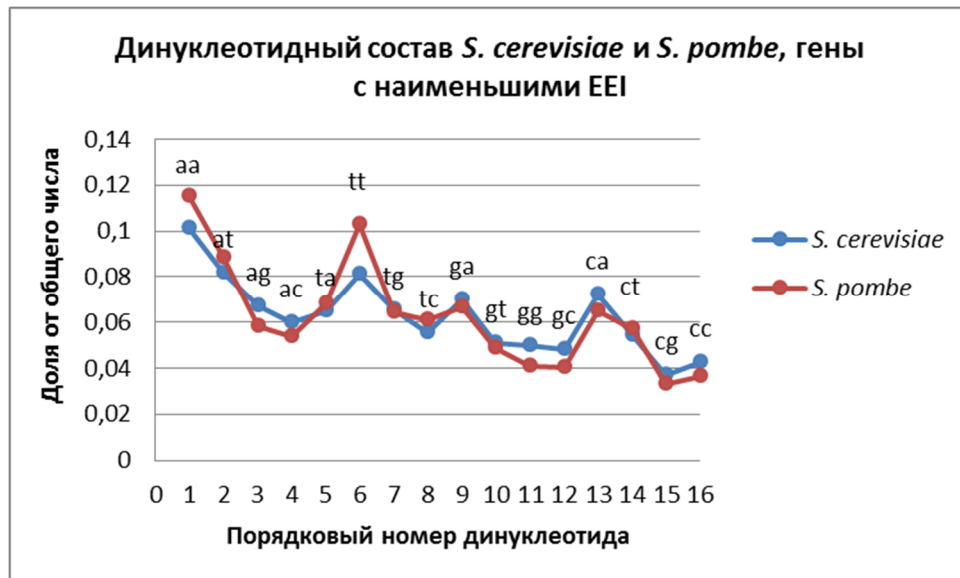
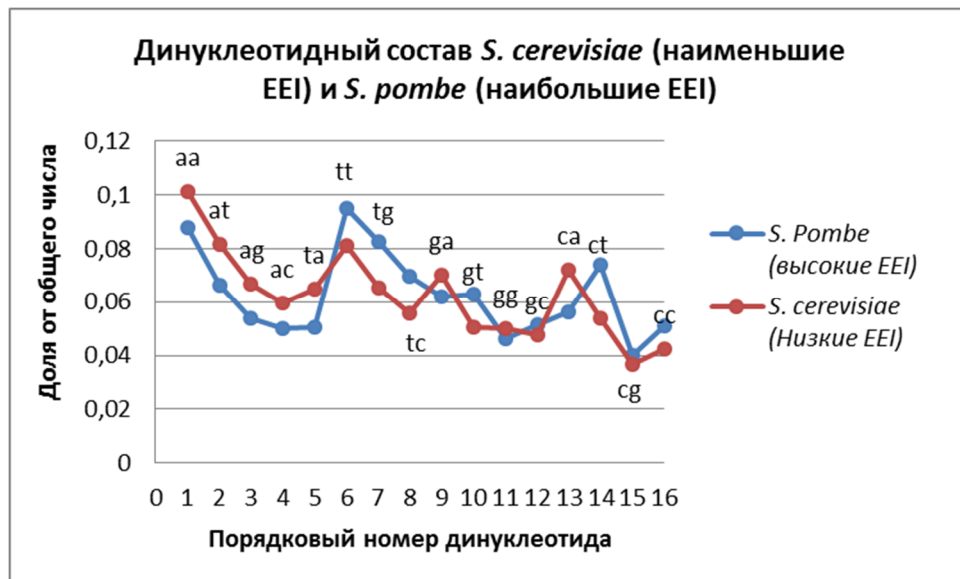


Рис. 71. Динуклеотидный состав 15% генов *S. cerevisiae* и *S. pombe* с наименьшими значениями EEI.

**Таблица 9.** Коэффициенты корреляции между динуклеотидными составами генов *S. cerevisiae* и *S. pombe* (красный фон – корреляция между аналогичными группами генов разных организмов, зеленый – корреляция между разными группами генов одного организма).

		Все гены		Высоко-экспрессирующиеся		Низко-экспрессирующиеся	
		<i>S.cerevisiae</i>	<i>S.pombe</i>	<i>S.cerevisiae</i>	<i>S.pombe</i>	<i>S.cerevisiae</i>	<i>S.pombe</i>
Все гены	<i>S.cerevisiae</i>	1,00	0,94	0,95	0,72	0,99	0,96
	<i>S.pombe</i>		1,00	0,93	0,88	0,92	0,99
Высоко-экспрессирующиеся	<i>S.cerevisiae</i>			1,00	0,85	0,92	0,91
	<i>S.pombe</i>				1,00	0,68	0,80
Низко-экспрессирующиеся	<i>S.cerevisiae</i>					1,00	0,96
	<i>S.pombe</i>						1,00

Самое низкое значение коэффициента корреляции  $r = 0,68$  ( $P > 99\%$ ) наблюдается между низкоэкспрессирующимися генами *S. cerevisiae* и высокоэкспрессирующимися генами *S. pombe* (Рисунок 72). Это те самые наборы генов, у которых наблюдается самое низкое значение среднего ПФН в кодирующей части генов (Рисунки 61-62).



**Рис. 72.** Динуклеотидный состав 15% генов *S. cerevisiae* с наименьшими значениями EEI и 15% генов *S. pombe* с наибольшими значениями EEI.

Частоты нуклеотидов «А» и «Т» для данных генов также различны (Рисунки 73-74). В низкоэкспрессирующихся генах *S. cerevisiae* чаще встречается нуклеотид «А», а в

высокоэкспрессирующихся генах *S. pombe* – нуклеотид «Т». Стоит отметить, что подобное распределение, как на Рисунке 74, больше ни в одной из групп генов не встречается.



Рис. 73. Частоты нуклеотидов в 15% генов *S. cerevisiae* с наименьшими значениями EEI.

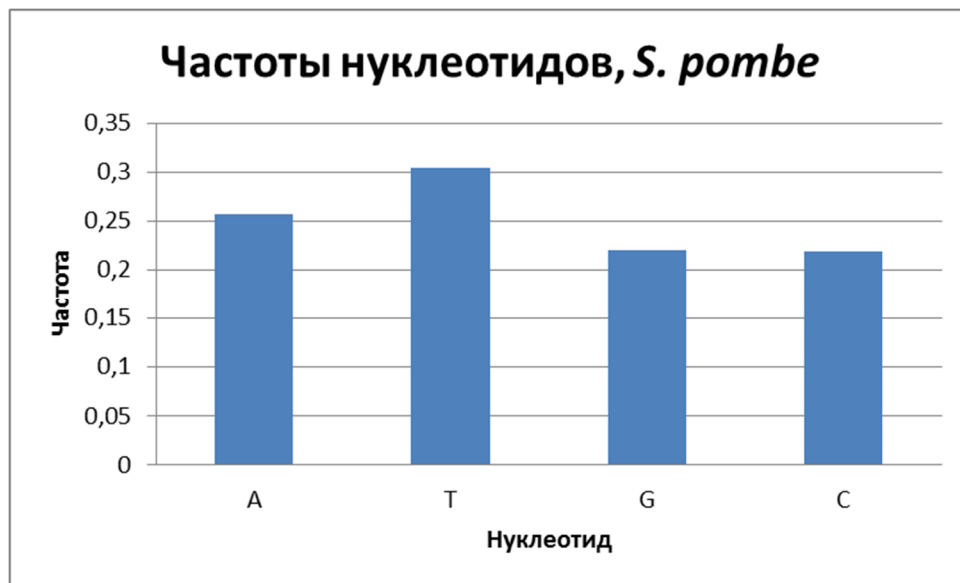


Рис. 74. Частоты нуклеотидов в 15% генов *S. pombe* с наибольшими значениями EEI.

### 3.5.7. Результаты исследования возможных причин различия видов корреляций между ПФН и EEI у *S. cerevisiae* и *S. pombe*

Ниже приведены наибольшие различия, выявленные при сравнении разных групп генов *S. cerevisiae* и *S. pombe*.

1) Высокоэкспрессирующиеся гены *S. pombe* содержат в своем составе больше нуклеотидов «Т», чем «А», в отличие от низкоэкспрессирующихся генов *S. cerevisiae* (Рисунки 73-74);

2) Высокоэкспрессирующиеся гены *S. pombe* обладают самым высоким GC-составом (Рисунок 65);

3) Высокоэкспрессирующиеся гены *S. pombe* наиболее сильно отличаются по динуклеотидному составу от низкоэкспрессирующихся генов *S. cerevisiae* (Рисунок 72, Таблица 9);

4) 33% генов *S. pombe*, имеющих такие же идентификаторы, как 15% генов *S. cerevisiae* с наибольшими значениями EEI, не попадают в группу 15% высокоэкспрессирующихся генов данного организма (Рисунки 67-68);

Таким образом, можно сделать вывод, что различия в значениях коэффициентов корреляции для кодирующих районов генов *S. cerevisiae* и *S. pombe* (Рисунки 55-58) связано именно с различием в их первичной структуре. Отбор кодонов в генах данных организмов идет по-разному. Это согласуется с данными из раздела 1.2.1.3. главы «Обзор литературы» о том, что у разных видов дрожжей ортологичные гены по-разному адаптируют свой кодонный состав к похожим пулам тРНК в зависимости от функциональной значимости этих генов для жизни каждого конкретного организма.

Различие в первичной структуре генов *S. cerevisiae* и *S. pombe* приводит к тому, что в кодирующей части генов наименьший нуклеосомный потенциал наблюдается именно в низкоэкспрессирующихся генах *S. cerevisiae* и высокоэкспрессирующихся генах *S. pombe* (Рисунки 61-62). Как было показано выше, есть значительная корреляция между значением нуклеосомного потенциала и коэффициентом корреляции между ПФН и EEI. Видимо, первичная структура низкоэкспрессирующихся генов у *S. cerevisiae* оптимизирована таким образом, чтобы при замедлении процесса транскрипции из-за образования нуклеосомного комплекса в районе, где у большинства генов вероятность такого образования низка, ускорялся процесс элонгации трансляции, для компенсации данной задержки. Иначе эффективность экспрессии гена может значительно снизиться, если будет замедлен и процесс транскрипции и трансляции. У *S. pombe* – наоборот, у высокоэкспрессирующихся генов должны быть высокоэффективны как процесс транскрипции, так и трансляции, поэтому при снижении вероятности образования нуклеосомного комплекса, также ускоряется элонгация трансляции.

## ЗАКЛЮЧЕНИЕ

На основе алгоритма расчета индекса EEI была реализована программа EtoE. Данная программа позволяет считывать из файла в gbk-формате белок-кодирующие последовательности (CDS) в геноме организма и рассчитывать для них значения индекса EEI. Также программа определяет, к какой из пяти групп по эволюционной оптимизации процесса элонгации трансляции относится данный организм.

Применение описанной программы позволило проанализировать геномы 2582 бактерий, 165 архей и 24 одноклеточных эукариот. Использование генов рибосомных белков как маркеров высокоэкспрессирующихся генов позволило показать высокую корреляцию индекса EEI с уровнем экспрессии данных генов.

Был проведен биоинформатический анализ геномов 62 штаммов, принадлежащих к 27 видам *Mycoplasma*. Установлено, что эволюционный отбор по эффективности трансляции генов у данных организмов шел на минимизацию количества потенциальных вторичных структур в мРНК, а не на оптимизацию кодонного состава. Получена новая информация о количестве локальных инвертированных повторов в генах разных штаммов *Mycoplasma*. Была обнаружена группа видов микоплазм (*C. M. haemolamae*, *C. M. haemominutum*, *M. haemocanis*, *M. haemofelis*, *M. pneumoniae*, *M. suis*) с пониженным содержанием в генах локальных инвертированных повторов. Было показано, что почти все они (кроме *M. pneumoniae*) относятся к одной филогенетической группе гемоплазм и характеризуются обитанием на поверхности или внутри эритроцитов. Также при построении профилей распределения локальных инвертированных повторов в районах старт- и стоп-кодона трансляции у *M. haemofelis* обнаружен не характерный для остальных микоплазм пик в районе старт-кодона. Пик свидетельствует о повышенной стабильности шпилек в данном районе гена. Выдвинуто предположение, что это может быть связано с наличием у *M. haemofelis* альтернативного механизма регуляции трансляции, отличного от механизма у других видов микоплазм.

При анализе реконструированных предковых последовательностей генов архей удалось показать связь между наиболее сильными изменениями в значениях их индекса EEI и сменой условий среды обитания данных организмов.

При анализе геномов *S. cerevisiae* и *S. pombe* была подтверждена предложенная гипотеза о согласованности оптимизации процессов транскрипции и трансляции. Показана взаимосвязь между локализацией нуклеосом в промоторных районах и индексом эффективности элонгации. Отбор может идти как на уменьшение ПФН для высокоэкспрессирующихся генов, так и на увеличение ПФН для низкоэкспрессирующихся.

Для генов *S. pombe* с высоким уровнем трансляции (по EEI) наблюдаются районы с достоверной отрицательной корреляцией индекса эффективности элонгации трансляции и потенциала формирования нуклеосом в 5'-районе выше старта транскрипции. Тогда как для генов *S. cerevisiae* с низким уровнем трансляции наблюдаются районы с достоверной положительной корреляцией.

Для дрожжей вида *S. cerevisiae* в кодирующей части генов с низким уровнем трансляции наблюдается высокая достоверная положительная корреляция между индексом эффективности элонгации трансляции и нуклеосомным потенциалом. Тогда как для дрожжей вида *S. pombe* в кодирующей части генов с высоким уровнем трансляции в этом же районе наблюдается высокая отрицательная корреляция нуклеосомного потенциала и индекса эффективности элонгации трансляции.

Дополнительно проведенный анализ выявил различия в первичной структуре генов исследуемых дрожжей. Отбор кодонов в генах данных организмов идет по-разному. Это приводит к тому, что в кодирующей части генов наименьший нуклеосомный потенциал наблюдается именно в низкоэкспрессирующихся генах *S. cerevisiae* и высокоэкспрессирующихся генах *S. pombe*. Видимо, первичная структура низкоэкспрессирующихся генов у *S. cerevisiae* оптимизирована таким образом, чтобы при замедлении процесса транскрипции из-за образования нуклеосомного комплекса в районе, где у большинства генов вероятность такого образования низка, ускорился процесс элонгации трансляции, для компенсации данной задержки. Иначе эффективность экспрессии гена может значительно снизиться, если будет замедлен и процесс транскрипции и трансляции. А у *S. pombe* – наоборот, у высокоэкспрессирующихся генов должны быть высокоэффективны как процесс транскрипции, так и трансляции, поэтому при снижении вероятности образования нуклеосомного комплекса, также ускоряется элонгация трансляции.



## ВЫВОДЫ

- 1) Создано веб-приложение EloE, реализующее самообучающийся алгоритм расчета индекса эффективности элонгации EEI, которое обеспечивает достоверный высокопроизводительный анализ эффективности элонгации трансляции генов одноклеточных организмов и позволяет обрабатывать до нескольких тысяч геномов за один запуск.
- 2) С помощью программы EloE проведена классификация 2771 генома различных одноклеточных организмов (бактерии, археи, эукариоты) по пяти типам эволюционной оптимизации процесса элонгации трансляции. Показано различие между бактериями и археями по предпочтительному типу оптимизации: у большинства проанализированных бактерий (45,35% от 2582) основную роль в определении эффективности элонгации трансляции играет только кодонный состав генов; у большинства архей (58,79% от 165) – кодонный состав генов и количество локальных инвертированных повторов в мРНК.
- 3) Показано наличие достоверной отрицательной корреляции ( $r = -0,85$ ;  $P < 1,97 \times 10^{-18}$ ) между GC-составом генома и степенью эволюционной оптимизации первичной структуры генов *Mycoplasma* для повышения эффективности элонгации трансляции.
- 4) Показано, что у семи видов *Mycoplasma* (*C. M. haemolamae*, *M. haemocanis*, *M. wenyonii*, *M. haemofelis*, *M. pneumonia*, *C. M. haemominutum*, *M. suis*), в отличие от 20 других, в процессе эволюции прошла массовая минимизация количества локальных совершенных инвертированных повторов (потенциальных шпилек) в мРНК.
- 5) Анализ профилей стабильности потенциальных вторичных структур в мРНК у *Mycoplasma* показал, что *M. haemofelis* радикально отличается от остальных видов микоплазм наличием более стабильных потенциальных вторичных структур в мРНК в районе старт-кодона трансляции.
- 6) Показано наличие достоверной отрицательной корреляции ( $r = -0,40$ ;  $P < 8,65 \times 10^{-8}$ ) между GC-составом генома и степенью эволюционной оптимизации первичной структуры генов архей для повышения эффективности элонгации трансляции.
- 7) Показано наличие значимой корреляции между потенциалом формирования нуклеосом и индексом эффективности элонгации трансляции генов *S. cerevisiae* и *S. pombe*.

## СПИСОК ЛИТЕРАТУРЫ

1. Владимиров Н.В., Лихошвай В.А., Матушкин Ю.Г. Корреляция частот кодонов и потенциальных вторичных структур с эффективностью трансляции мРНК в одноклеточных организмах // Молекулярная биология. – 2007. – Т. 41. – № 5. – С. 926–933.
2. Лихошвай В.А., Матушкин Ю.Г. Предсказание эффективности экспрессии генов по их нуклеотидному составу // Молекулярная биология. – 2000. – Т. 34. – № 3. – С. 406–412.
3. Amo T., Paje M.L.F., Inagaki A., Ezaki S., Atomi H., Imanaka T. *Pyrobaculum calidifontis* sp. nov., a novel hyperthermophilic archaeon that grows in atmospheric air // Archaea. – 2002. – Vol. 1. – No. 2. – P. 113–121.
4. Andersson S.G.E., Kurland C.G. Codon preferences in free-living microorganisms // Microbiol. – 1990. – Rev. 54. – P. 198–210.
5. Antunes A., Taborda M., Huber R., Moissl C., Nobre M.F., da Costa M.S. *Halorhabdus tiamatea* sp. nov., a non-pigmented, extremely halophilic archaeon from a deep-sea, hypersaline anoxic basin of the Red Sea, and emended description of the genus *Halorhabdus* // International Journal of Systematic and Evolutionary Microbiology. – 2008. – Vol. 58. – No. 1. – P. 215–220.
6. Arab H., Völker H., Thomm M. *Thermococcus aegaicus* sp. nov. and *Staphylothermus hellenicus* sp. nov., two novel hyperthermophilic archaea isolated from geothermally heated vents off Palaeochori Bay, Milos, Greece // International journal of systematic and evolutionary microbiology. – 2000. – Vol. 50. – No. 6. – P. 2101–2108.
7. Arava Y., Wang Y., Storey J.D., Liu C.L., Brown P.O., Herschlag D. Genome-wide analysis of mRNA translation profiles in *Saccharomyces cerevisiae* // Proceedings of the National Academy of Sciences. – 2003. – Vol. 100. – No. 7. – P. 3889–3894.
8. Bennetzen J.L., Hall B.D. Codon selection in Yeast // J. Biol. Chem. – 1982. – Vol. 257. – P. 3026–3031.
9. Birrien J.L., Zeng X., Jebbar M., Cambon-Bonavita M.A., Quérellou J., Oger P., Bienvenu N., Xiao X., Prieur, D. *Pyrococcus yayanosii* sp. nov., an obligate piezophilic hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent // International journal of systematic and evolutionary microbiology. – 2011. – Vol. 61. – No. 12. – P. 2827–2881.
10. Blöchl E., Rachel R., Burggraf S., Hafenbradl D., Jannasch H.W., Stetter K.O. *Pyrolobus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113°C // Extremophiles. – 1997. – Vol. 1. – No. 1. – P. 14–21.

11. Bonch-Osmolovskaya E.A., Miroschnichenko M.L., Kostrikina N.A., Chernych N.A., Zavarzin G.A. *Thermoproteus uzoniensis* sp. nov., a new extremely thermophilic archaeobacterium from Kamchatka continental hot springs // Archives of microbiology. – 1990. – Vol. 154. – No. 6. – P. 556–559.
12. Bonitz S.G., Berlani R., Coruzzi G., Li M., Macino G., Nobrega F.G., Tzagoloff A. Codon recognition rules in yeast mitochondria // Proceedings of the National Academy of Sciences. – 1980. – Vol. 77. – No. 6. – P. 3167–3170.
13. Borer P.N. Dengler B., Tinoco I.Jr. Stability of ribonucleic acid double-stranded helices // Journal of molecular biology. – 1974. – Vol. 86. – No. 4. – P. 843–853.
14. Borges N., Matsumi R., Imanaka T., Atomi H., Santos H. *Thermococcus kodakarensis* mutants deficient in di-myo-inositol phosphate use aspartate to cope with heat stress // Journal of bacteriology. – 2010. – Vol. 192. – No. 1. – P. 191–197.
15. Bulmer M. Coevolution of codon usage and transfer RNA abundance // Nature. – 1987. – Vol. 325. – No. 6106. – P. 728–730.
16. Burns D.G., Janssen P.H., Itoh T., Kamekura M., Li Z., Jensen G., Rodríguez-Valera F., Bolhuis H., Dyall-Smith M.L. *Haloquadratum walsbyi* gen. nov., sp. nov., the square haloarchaeon of Walsby, isolated from saltern crystallizers in Australia and Spain // International Journal of Systematic and Evolutionary Microbiology. – 2007. – Vol. 57. – No. 2. – P. 387–392.
17. Burns D.G. Janssen P.H., Itoh T., Minegishi H., Usami R., Kamekura M., Dyall-Smith M.L. *Natronomonas moolapensis* sp. nov., non-alkaliphilic isolates recovered from a solar saltern crystallizer pond, and emended description of the genus *Natronomonas* // International journal of systematic and evolutionary microbiology. – 2010. – Vol. 60. – No. 5. – P. 1173–1176.
18. Cadillo-Quiroz H., Bräuer S.L., Goodson N., Yavitt J. B., Zinder S.H. *Methanobacterium paludis* sp. nov. and a novel strain of *Methanobacterium lacus* isolated from northern peatlands // International journal of systematic and evolutionary microbiology. – 2014. – Vol. 64. – No. Pt 5. – P. 1473–1480.
19. Cannarozzi G., Schraudolph N.N., Faty M., von Rohr P., Friberg M.T., Roth A.C., Gonnet P., Gonnet G., Barrall Y. A role for codon order in translation dynamics // Cell. – 2010. – Vol. 141. – No. 2. – P. 355–367.
20. Carbone A., Kepes F., Zinovyev A. Codon bias signatures, organization of microorganisms in codon space, and lifestyle // Molecular biology and evolution. – 2005. – Vol. 22. – No. 3. – P. 547–561.
21. Castillo A. M. Gutiérrez M.C., Kamekura M., Xue Y., Ma Y., Cowan D.A., Jones B.E., Grant W.D., Ventosa A. *Halovivax ruber* sp. nov., an extremely halophilic archaeon isolated from

- Lake Xilinhot, Inner Mongolia, China // International journal of systematic and evolutionary microbiology. – 2007. – Vol. 57. – No. 5. – P. 1024–1027.
22. Chan P. P., Cozen A. E., Lowe T. M. Reclassification of *Thermoproteus neutrophilus* Stetter and Zillig 1989 as *Pyrobaculum neutrophilum* comb. nov. based on phylogenetic analysis // International journal of systematic and evolutionary microbiology. – 2013. – Vol. 63. – No. Pt 2. – P. 751–754.
  23. Ciandrini L., Stansfield I., Romano M.C. Ribosome traffic on mRNAs maps to gene ontology: genome-wide quantification of translation initiation rates and polysome size regulation // PLoS computational biology. – 2013. – Vol. 9. – No. 1. – P. e1002866.
  24. Collier A.M., Clyde W.A.Jr. Relationships between *Mycoplasma pneumoniae* and human respiratory epithelium // Infect. Immun. – 1971. – Vol. 3. – No. 5. – P. 694–701.
  25. Comeron J. M., Aguadé M. An evaluation of measures of synonymous codon usage bias // Journal of molecular evolution. – 1998. – Vol. 47. – No. 3. – P. 268–274.
  26. Dam E.B., Pleij C.W., Bosch L. RNA pseudoknots: translational frameshifting and readthrough on viral RNAs // Virus Genes. – 1990. – Vol. 4. – P. 121–136.
  27. De Rosa M., Gambacorta A., Bu'Lock J. D. Extremely thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius* // Journal of general microbiology. – 1975. – Vol. 86. – No. 1. – P. 156–164.
  28. Delisi C., Crothers D.M. Prediction of RNA secondary structure // Proceedings of the National Academy of Sciences. – 1971. – Vol. 68. – No. 11. – P. 2682–2685.
  29. Ding Y., Tang Y., Kwok C.K., Zhang Y., Bevilacqua P.C., Assmann S.M. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features // Nature. – 2014. – Vol. 505. – No. 7485. – P. 696–700.
  30. Dittmar K.A., Goodenbour J.M., Pan T. Tissue-specific differences in human transfer RNA expression // PLoS genetics. – 2006. – Vol. 2. – No. 12. – P. e221.
  31. Diwa A., Bricker A.L., Jain C., Belasco J.G. An evolutionarily conserved RNA stem-loop functions as a sensor that directs feedback regulation of RNase E gene expression // Genes & development. – 2000. – Vol. 14. – No. 10. – P. 1249–1260.
  32. Dopson M., Baker-Austin C., Hind A., Bowman J.P., Bond P.L. Characterization of *Ferroplasma* isolates and *Ferroplasma acidarmanus* sp. nov., extreme acidophiles from acid mine drainage and industrial bioleaching environments // Applied and environmental microbiology. – 2004. – Vol. 70. – No. 4. – P. 2079–2088.
  33. Doty P., Boedtke H., Fresco J.R., Haselkorn R., Litt M. Secondary structure in ribonucleic acids // Proceedings of the National Academy of Sciences of the United States of America. – 1959. – Vol. 45. – No. 4. – P. 482.

34. Eck S., Stephan W. Determining the relationship of gene expression and global mRNA stability in *Drosophila melanogaster* and *Escherichia coli* using linear models // *Gene*. – 2008. – Vol. 424. – No. 1. – P. 102–107.
35. Elkins J.G., Podar M., Graham D.E., Makarova K.S., Wolf Y., Randau L., Hedlund B.P., Brochier-Armanet C., Kunin V., Anderson I., Lapidus A., Goltsman E., Barry K., Koonin E.V., Hugenholtz P., Kyrpides N., Wanner G., Richardson P., Keller M., Stetter K.O. A korarchaeal genome reveals insights into the evolution of the Archaea // *Proceedings of the National Academy of Sciences*. – 2008. – Vol. 105. – No. 23. – P. 8102–8107.
36. Erauso G., Reysenbach A.L., Godfroy A., Meunier J.R., Crump B., Partensky F., Baross J.A., Marteinsson V., Barbier G., Pace N.R., Prieur D. *Pyrococcus abyssi* sp. nov., a new hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent // *Archives of Microbiology*. – 1993. – Vol. 160. – No. 5. – P. 338–349.
37. Ermolaeva M.D. Synonymous codon usage in bacteria // *Current issues in molecular biology*. – 2001. – Vol. 3. – No. 4. – P. 91–97.
38. Fan H., Xue Y., Ma Y., Ventosa A., Grant W.D.. *Halorubrum tibetense* sp. nov., a novel haloalkaliphilic archaeon from Lake Zabuye in Tibet, China // *International journal of systematic and evolutionary microbiology*. – 2004. – Vol. 54. – No. 4. – P. 1213–1216.
39. Fredrick K., Ibba M. How the sequence of a gene can tune its translation // *Cell*. – 2010. – Vol. 141. – No. 2. – P. 227–229.
40. Fiala G., Stetter K. O. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaeobacteria growing optimally at 100°C // *Archives of Microbiology*. – 1986. – Vol. 145. – No. 1. – P. 56–61.
41. Fiala G., Stetter K.O., Jannasch H.W., Langworthy T.A., Madon J. *Staphylothermus marinus* sp. nov. represents a novel genus of extremely thermophilic submarine heterotrophic archaeobacteria growing up to 98°C // *Systematic and Applied Microbiology*. – 1986. – Vol. 8. – No. 1. – P. 106–113.
42. Fresco J.R., Alberts B.M., Doty P. Some molecular details of the secondary structure of ribonucleic acid // *Nature*. – 1960. – Vol. 188. – P. 98–101.
43. Fuglsang A. Codon optimizer: a freeware tool for codon optimization // *Protein expression and purification*. – 2003. – Vol. 31. – No. 2. – P. 247–249.
44. Ganem D., Miller J. H., Files J. G., Platt T., Weber K. Reinitiation of a *lac* repressor fragment at a codon other than AUG // *Proceedings of the National Academy of Sciences*. – 1973. – Vol. 70. – No. 11. – P. 3165–3169.

45. Gaspar P., Moura G., Santos M.A.S., Oliveira J.L. mRNA secondary structure optimization using a correlated stem-loop prediction // *Nucleic acids research*. – 2013. – Vol. 41. – No. 6. – P. e73–e73.
46. González J. M., Masuchi Y., Robb F.T., Ammerman J.W., Maeder D.L., Yanagibayashi M., Tamaoka J., Kato C. *Pyrococcus horikoshii* sp. nov., a hyperthermophilic archaeon isolated from a hydrothermal vent at the Okinawa Trough // *Extremophiles*. – 1998. – Vol. 2. – No. 2. – P. 123–130.
47. Goh W.S., Orlov Y., Li J., Clarke N.D. Blurring of high-resolution data shows that the effect of intrinsic nucleosome occupancy on transcription factor binding is mostly regional, not local // *PLoS Comput Biol*. – 2010. – Vol. 6. – No. 1. – P. e1000649.
48. Gouy M., Gautier C. Codon usage in bacteria: correlation with gene expressivity // *Nucleic Acids Res*. – 1982. – Vol. 10. – P. 7055–7070.
49. Grantham R. Working of the genetic code // *Trends in Biochemical Sciences*. – 1980. – Vol. 5. – No. 12. – P. 327–331.
50. Grantham R., Gautier C., Gouy M., Mercier R., Pavé A. Codon catalog usage and the genome hypothesis // *Nucleic Acids Res*. – 1980. – Vol. 8. – P. r49–r62.
51. Grantham R., Gautier C., Gouy M., Jacobzone M., Mercier R. Codon catalog usage is a genome strategy modulated for gene expressivity // *Nucleic Acids Research*. – 1981. – Vol. 9. – No. 1. – P. 213–213.
52. Gribskov M., Devereux J., Burgess R.R. The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression // *Nucleic acids research*. – 1984. – Vol. 12. – No. 1 part 2. – P. 539–549.
53. Grote A., Hiller K., Scheer M., Münch R., Nörtemann B., Hempel D.C., Jahn D. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host // *Nucleic acids research*. – 2005. – Vol. 33. – No. suppl. 2. – P. W526–W531.
54. Gruber C., Legat A., Pfaffenhuemer M., Radax C., Weidler G., Busse H.J., Stan-Lotter H. *Halobacterium noricense* sp. nov., an archaeal isolate from a bore core of an alpine Permian salt deposit, classification of *Halobacterium* sp. NRC-1 as a strain of *H. salinarum* and emended description of *H. salinarum* // *Extremophiles*. – 2004. – Vol. 8. – No. 6. – P. 431–439.
55. Gu W., Zhou T., Wilke C.O. A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes // *PLoS computational biology*. – 2010. – Vol. 6. – No. 2. – P. e1000664.
56. Guisez Y., Robbens J., Remaut E., Fiers W. Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon

- usage: A hypothesis // *Journal of theoretical biology*. – 1993. – Vol. 162. – No. 2. – P. 243–252.
57. Gumerov V.M., Mardanov A.V., Beletsky A.V., Prokofeva M.I., Bonch-Osmolovskaya E.A., Ravin N.V., Skryabin K.G. Complete genome sequence of “*Vulcanisaeta moutnovskia*” strain 768-28, a novel member of the hyperthermophilic crenarchaeal genus *Vulcanisaeta* // *Journal of bacteriology*. – 2011. – Vol. 193. – No. 9. – P. 2355–2356.
58. Gutiérrez M.C., Castillo A.M., Kamekura M., Xue Y., Ma Y., Cowan D.A., Jones B.E., Grant W.D., Ventosa A. *Halopiger xanaduensis* gen. nov., sp. nov., an extremely halophilic archaeon isolated from saline Lake Shangmatale in Inner Mongolia, China // *International journal of systematic and evolutionary microbiology*. – 2007. – Vol. 57. – No. 7. – P. 1402–1407.
59. Hafenbradl D., Keller M., Dirmeier R., Rachel R., Roßnagel P., Burggraf S., Huber H., Stetter K.O. *Ferroglobus placidus* gen. nov., sp. nov., a novel hyperthermophilic archaeum that oxidizes Fe<sup>2+</sup> at neutral pH under anoxic conditions // *Archives of Microbiology*. – 1996. – Vol. 166. – No. 5. – P. 308–314.
60. Hall M.N. Gabay J., Débarbouillé M., Schwartz M. A role for mRNA secondary structure in the control of translation initiation. – 1982. – P. 616–618.
61. Hershberg R., Petrov D. A. Selection on codon bias // *Annual review of genetics*. – 2008. – Vol. 42. – P. 287–299.
62. Hoede C., Denamur E., Tenaillon O. Selection acts on DNA secondary structures to decrease transcriptional mutagenesis // *PLoS genetics*. – 2006. – Vol. 2. – No. 11. – P. e176.
63. Hofacker I. L. Vienna RNA secondary structure server // *Nucleic acids research*. – 2003. – Vol. 31. – No. 13. – P. 3429–3431.
64. Huber G., Spinnler C., Gambacorta A., Stetter, K.O. *Metallosphaera sedula* gen, and sp. nov. represents a new genus of aerobic, metal-mobilizing, thermoacidophilic archaeobacteria // *Systematic and applied microbiology*. – 1989. – Vol. 12. – No. 1. – P. 38–47.
65. Huber H., Jannasch H., Rachel R., Fuchs T., Stetter, K.O. *Archaeoglobus veneficus* sp. nov., a novel facultative chemolithoautotrophic hyperthermophilic sulfite reducer, isolated from abyssal black smokers // *Systematic and Applied Microbiology*. – 1997. – Vol. 20. – No. 3. – P. 374–380.
66. Huber R., Dyba D., Huber H., Burggraf S., Rachel R. Sulfur-inhibited *Thermosphaera aggregans* sp. nov., a new genus of hyperthermophilic archaea isolated after its prediction from environmentally derived 16S rRNA sequences // *International journal of systematic bacteriology*. – 1998. – Vol. 48. – No. 1. – P. 31–38.
67. Huber H., Prangishvili D. Sulfolobales // *The Prokaryotes*. – Springer New York, 2006. – P. 23–51.

68. Ikemura T. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* system // *Journal of Molecular Biology*. – 1981. – Vol. 151. – P. 389–409.
69. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes: differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs // *Journal of molecular biology*. – 1982. – Vol. 158. – No. 4. – P. 573–597.
70. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms // *Mol. Biol. Evol.* – 1985. – Vol. 2. – P. 13–34.
71. Ikemura T., Ozeki H. Codon usage and transfer RNA contents: organism specific codon choice patterns in reference to the isoacceptor contents. // *Cold Spring Harbor Symposium Quantitative Biology*. – 1982. – Vol. 47. – P. 1087–1097.
72. Ingolia N.T., Ghaemmaghami S., Newman J.R.S., Weissman J.S. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling // *Science*. – 2009. – Vol. 324. – No. 5924. – P. 218–223.
73. Itoh T., Suzuki K.I., Sanchez P.C., Nakase T. *Caldivirga maquilingensis* gen. nov., sp. nov., a new genus of rod-shaped crenarchaeote isolated from a hot spring in the Philippines // *International journal of systematic bacteriology*. – 1999. – Vol. 49. – No. 3. – P. 1157–1163.
74. Itoh T., Suzuki K., Nakase T. *Vulcanisaeta distributa* gen. nov., sp. nov., and *Vulcanisaeta souniana* sp. nov., novel hyperthermophilic, rod-shaped crenarchaeotes isolated from hot springs in Japan // *International journal of systematic and evolutionary microbiology*. – 2002. – Vol. 52. – No. 4. – P. 1097–1104.
75. Itoh T., Suzuki K., Sanchez P.C., Nakase T. *Caldisphaera lagunensis* gen. nov., sp. nov., a novel thermoacidophilic crenarchaeote isolated from a hot spring at Mt Maquiling, Philippines // *International journal of systematic and evolutionary microbiology*. – 2003. – Vol. 53. – No. 4. – P. 1149–1154.
76. Jacks T., Madhani H.D., Masiarz F.R., Varmus H.E. Signals for ribosomal frameshifting in the Rous sarcoma virus *gag-pol* region // *Cell*. – 1988. – Vol. 55. – No. 3. – P. 447–458.
77. Jahn U., Summons R., Sturt H., Grosjean E., Huber H. Composition of the lipids of *Nanoarchaeum equitans* and their origin from its host *Ignicoccus* sp. strain KIN4/I // *Archives of microbiology*. – 2004. – Vol. 182. – No. 5. – P. 404–413.
78. Jahn U., Hohn M.J., Kronner M., Näther D.J., Burghardt T., Rachel R., Stetter K.O., Huber H. *Ignicoccus hospitalis* sp. nov., the host of ‘*Nanoarchaeum equitans*’ // *International journal of systematic and evolutionary microbiology*. – 2007. – Vol. 57. – No. 4. – P. 803–808.



79. von Jan M., Lapidus A., Del Rio T.G., Copeland A., Tice H., Cheng J.F., Lucas S., Chen F., Nolan M., Goodwin L., Han C., Pitluck S., Liolios K., Ivanova N., Mavromatis K., Ovchinnikova G., Chertkov O., Pati A., Chen A., Palaniappan K., Land M., Hauser L., Chang Y.-J., Jeffries C.D., Saunders E., Brettin T., Detter J.C., Chain P., Eichinger K., Huber H., Spring S., Rohde M., Göker M., Wirth R., Woyke T., Bristow J., Eisen J.A., Markowitz V., Hugenholtz P., Kyrpides N.C., Klenk H.P. Complete genome sequence of *Archaeoglobus profundus* type strain (AV18T) // Standards in genomic sciences. – 2010. – Vol. 2. – No. 3. – P. 327.
80. Jia M., Luo L. The relation between mRNA folding and protein structure // Biochemical and biophysical research communications. – 2006. – Vol. 343. – No. 1. – P. 177–182.
81. Jolivet E., L'Haridon S., Corre E., Forterre P., Prieur D. *Thermococcus gammatolerans* sp. nov., a hyperthermophilic archaeon from a deep-sea hydrothermal vent that resists ionizing radiation // International Journal of Systematic and Evolutionary Microbiology. – 2003. – Vol. 53. – No. 3. – P. 847–851.
82. Kaczanowska M., Rydén–Aulin M. Ribosome biogenesis and the translation process in *Escherichia coli* // Microbiology and Molecular Biology Reviews. – 2007. – Vol. 71. – No. 3. – P. 477–494.
83. Kaplan N., Moore I.K., Fondufe-Mittendorf Y., Gossett A.J., Tillo D., Field Y., LeProust E.M., Hughes T.R., Lieb J.D., Widom J., Segal E. The DNA-encoded nucleosome organization of a eukaryotic genome // Nature. – 2009. – Vol. 458. – No. 7236. – P. 362–366.
84. Katz L., Burge C.B. Widespread selection for local RNA secondary structure in coding regions of bacterial genes // Genome Research. – 2003. – Vol. 13. – No. 9. – P. 2042–2051.
85. Kertesz M., Wan Y., Mazor E., Rinn J.L., Nutter R.C., Chang H.Y., Segal E. Genome-wide measurement of RNA secondary structure in yeast // Nature. – 2010. – Vol. 467. – No. 7311. – P. 103–107.
86. Klenk H.P., Clayton R.A., Tomb J.F., White O., Nelson K.E., Ketchum K.A., Dodson R.J., Gwinn M., Hickey E.K., Peterson J.D., Richardson D.L., Kerlavage A.R., Graham D.E., Kyrpides N.C., Fleischmann R.D., Quackenbush J., Lee N.H., Sutton G.G., Gill S., Kirkness E.F., Dougherty B.A., McKenney K., Adams M.D., Loftus B., Peterson S., Reich C.I., McNeil L.K., Badger J.H., Glodek A., Zhou L., Overbeek R., Gocayne J.D., Weidman J.F., McDonald L., Utterback T., Cotton M.D., Spriggs T., Artiach P., Kaine B.P., Sykes S.M., Sadow P.W., D'Andrea K.P., Bowman C., Fujii C., Garland S.A., Mason T.M., Olsen G.J., Fraser C.M., Smith H.O., Woese C.R., Venter J. C. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus* // Nature. – 1997. – Vol. 390. – No. 6658. – P. 364–370.

87. Klionsky D.J., Skalnik D.G., Simoni R.D. Differential translation of the genes encoding the proton-translocating ATPase of *Escherichia coli* // Journal of Biological Chemistry. – 1986. – Vol. 261. – No. 18. – P. 8096–8099.
88. Kozak M. Influences of mRNA secondary structure on initiation by eukaryotic ribosomes // Proceedings of the National Academy of Sciences. – 1986. – Vol. 83. – No. 9. – P. 2850–2854.
89. Kozak M. Circumstances and mechanisms of inhibition of translation by secondary structure in eucaryotic mRNAs // Molecular and cellular biology. – 1989. – Vol. 9. – No. 11. – P. 5134–5142.
90. Kublanov I.V., Bidjieva S.K., Mardanov A.V., Bonch-Osmolovskaya E.A. *Desulfurococcus kamchatkensis* sp. nov., a novel hyperthermophilic protein-degrading archaeon isolated from a Kamchatka hot spring // International journal of systematic and evolutionary microbiology. – 2009. – Vol. 59. – No. 7. – P. 1743–1747.
91. Lantermann A. B., Straub T., Strålfors A., Yuan G. C., Ekwall K., Korber P. *Schizosaccharomyces pombe* genome-wide nucleosome mapping reveals positioning mechanisms distinct from those of *Saccharomyces cerevisiae* // Nature structural & molecular biology. – 2010. – Vol. 17. – No. 2. – P. 251–257.
92. Levitsky V. G. RECON: a program for prediction of nucleosome formation potential // Nucleic acids research. – 2004. – Vol. 32. – No. suppl 2. – P. W346–W349.
93. Li H., Luo L. The relation between codon usage, base correlation and gene expression level in *Escherichia coli* and Yeast // J. Theor. Biol. Vol. – 1996. – 181. – No. 2. – P. 111–124.
94. Likhoshvai V.A., Matushkin Yu.G. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy // FEBS Letters. – 2002. – Vol. 516. – P. 87–92.
95. Liu L.J., You X.Y., Guo X., Liu S.J., Jiang C.Y. *Metallosphaera cuprina* sp. nov., an acidothermophilic, metal-mobilizing archaeon // International journal of systematic and evolutionary microbiology. – 2011. – Vol. 61. – No. 10. – P. 2395–2400.
96. Lopinski J.D., Dinman J.D., Bruenn J.A. Kinetics of ribosomal pausing during programmed –1 translational frameshifting // Mol. Cell. Biol. – 2000. – Vol. 20. – P. 1095–1103.
97. Man O., Pilpel Y. Differential translation efficiency of orthologous genes is involved in phenotypic divergence of yeast species // Nature genetics. – 2007. – Vol. 39. – No. 3. – P. 415–421.
98. Mao Y. Liu H., Liu Y., Tao S. Deciphering the rules by which dynamics of mRNA secondary structure affect translation efficiency in *Saccharomyces cerevisiae* // Nucleic acids research. – 2014. – Vol. 42. – No. 8. – P. 4813–4822.

99. Mardanov A.V., Kochetkova T.V., Beletsky A.V., Bonch-Osmolovskaya E.A., Ravin N.V., Skryabin K.G. Complete genome sequence of the hyperthermophilic cellulolytic crenarchaeon “*Thermogladius cellulolyticus*” 1633 // *Journal of bacteriology*. – 2012. – Vol. 194. – No. 16. – P. 4446–4447.
100. Markham N. R., Zuker M. UNAFold // *Bioinformatics*. – Humana Press, 2008. – P. 3–31.
101. Matushkin Yu.G., Likhoshvai V.A., Kochetov A.V. Local secondary structure may be a critical characteristic influencing translation of unicellular organisms mRNA // *Bioinformatics of Genome Regulation and Structure*. – 2004. – P. 103–114.
102. McGenity T. J., Gemmell R. T., Grant W. D. Proposal of a new halobacterial genus *Natrinema* gen. nov., with two species *Natrinema pellirubrum* nom. nov. and *Natrinema pallidum* nom. nov // *International journal of systematic bacteriology*. – 1998. – Vol. 48. – No. 4. – P. 1187–1196.
103. McLachlan A. D., Staden R., Boswell D.R. A method for measuring the non-random bias of a codon usage table // *Nucleic acids research*. – 1984. – Vol. 12. – No. 24. – P. 9567–9575.
104. Meijer H., Thomas A. Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA // *Biochem. J*. – 2002. – Vol. 367. – P. 1–11.
105. Miroshnichenko M.L., Hippe H., Stackebrandt E., Kostrikina N.A., Chernyh N.A., Jeanthon C., Nazina T.N., Belyaev S.S., Bonch-Osmolovskaya E.A. Isolation and characterization of *Thermococcus sibiricus* sp. nov. from a Western Siberia high-temperature oil reservoir // *Extremophiles*. – 2001. – Vol. 5. – No. 2. – P. 85–91.
106. Montalvo-Rodriguez R., Vreeland R.H., Oren A., Kessel M., Betancourt C., López-Garriga J. *Halogeometricum borinquense* gen. nov., sp. nov., a novel halophilic archaeon from Puerto Rico // *International journal of systematic bacteriology*. – 1998. – Vol. 48. – No. 4. – P. 1305–1312.
107. Nakagawa S., Takai K., Horikoshi K., Sako Y. *Aeropyrum camini* sp. nov., a strictly aerobic, hyperthermophilic archaeon from a deep-sea hydrothermal vent chimney // *International journal of systematic and evolutionary microbiology*. – 2004. – Vol. 54. – No. 2. – P. 329–335.
108. Naya H., Romero H., Carels N., Zavala A., Musto H. Translational selection shapes codon usage in the GC-rich genome of *Chlamydomonas reinhardtii* // *FEBS letters*. – 2001. – Vol. 501. – No. 2. – P. 127–130.
109. Neuner A., Jannasch H.W., Belkin S., Stetter K.O. *Thermococcus litoralis* sp. nov.: a new species of extremely thermophilic marine archaeobacteria // *Archives of microbiology*. – 1990. – Vol. 153. – No. 2. – P. 205–207.
110. Nichols B.P., Blumenberg M., Yanofsky C. Comparison of the nucleotide sequence of *trpA* and sequences immediately beyond the *trp* operon of *Klebsiella aerogenes*, *Salmonella*

- typhimurium* and *Escherichia coli* // Nucleic acids research. – 1981. – Vol. 9. – No. 7. – P. 1743–1756.
111. Niederberger T.D., Götz D.K., McDonald I.R., Ronimus R.S., Morgan H.W. *Ignisphaera aggregans* gen. nov., sp. nov., a novel hyperthermophilic crenarchaeote isolated from hot springs in Rotorua and Tokaanu, New Zealand // International journal of systematic and evolutionary microbiology. – 2006. – Vol. 56. – No. 5. – P. 965–971.
  112. Nussinov R., Pieczenik G., Griggs J. R., Kleitman D.J. Algorithms for loop matchings // SIAM Journal on Applied mathematics. – 1978. – Vol. 35. – No. 1. – P. 68–82.
  113. Nussinov R., Jacobson A.B. Fast algorithm for predicting the secondary structure of single-stranded RNA // Proceedings of the National Academy of Sciences. – 1980. – Vol. 77. – No. 11. – P. 6309–6313.
  114. O'Neill P.K., Or M., Erill I. scnRCA: A Novel method to detect consistent patterns of translational selection in mutationally-biased genomes // PloS one. – 2013. – Vol. 8. – No. 10. – P. e76177.
  115. Oren A., Ginzburg M., Ginzburg B.Z., Hochstein L.I., Volcani B.E. *Haloarcula marismortui* (Volcani) sp. nov., nom. rev., an extremely halophilic bacterium from the Dead Sea // International Journal of Systematic Bacteriology. – 1990. – Vol. 40. – No. 2. – P. 209–210.
  116. Oren A., Elevi R., Watanabe S., Ihara K., Corcelli A. *Halomicrobium mukohataei* gen. nov., comb. nov., and emended description of *Halomicrobium mukohataei* // International journal of systematic and evolutionary microbiology. – 2002. – Vol. 52. – No. 5. – P. 1831–1835.
  117. Osada Y., Saito R., Tomita M. Analysis of base-pairing potentials between 16S rRNA and 5'UTR for translation initiation in various prokaryotes // Bioinformatics. – 1999. – Vol. 15. – No. 7. – P. 578–581.
  118. Ouyang Z., Snyder M.P., Chang H.Y. SeqFold: Genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data // Genome research. – 2013. – Vol. 23. – No. 2. – P. 377–387.
  119. Perevalova A.A., Svetlichny V.A., Kublanov I.V., Chernyh N.A., Kostrikina N.A., Tourova T.P., Kuznetsov B. B., Bonch-Osmolovskaya E.A. *Desulfurococcus fermentans* sp. nov., a novel hyperthermophilic archaeon from a Kamchatka hot spring, and emended description of the genus *Desulfurococcus* // International journal of systematic and evolutionary microbiology. – 2005. – Vol. 55. – No. 3. – P. 995–999.
  120. Perevalova A.A., Bidzhieva S.K., Kublanov I.V., Hinrichs K.U., Liu X.L., Mardanov A.V., Lebedinsky A.V., Bonch-Osmolovskaya E.A. *Fervidicoccus fontis* gen. nov., sp. nov., an anaerobic, thermophilic crenarchaeote from terrestrial hot springs, and proposal of

- Fervidicoccaceae* fam. nov. and *Fervidicoccales* ord. nov // International journal of systematic and evolutionary microbiology. – 2010. – Vol. 60. – No. 9. – P. 2082–2088.
121. Peters I.R., Helps C.R., McAuliffe L., Neimark H., Lappin M.R., Gruffydd-Jones T.J., Day M.J., Hoelzle L.E., Willi B., Meli M., Hofmann-Lehmann R., Tasker S. RNase P RNA gene (*rnpB*) phylogeny of Hemoplasmas and other *Mycoplasma* species // J. Clin. Microbiol. – 2008. – Vol. 46. – No. 5. – 1873–1877.
  122. Platt T., Weber K., Ganem D., Miller J. H. Translational restarts: AUG reinitiation of a lac repressor fragment // Proceedings of the National Academy of Sciences. – 1972. – Vol. 69. – No. 4. – P. 897–901.
  123. Powell D.A., Hu P.C., Wilson M., Collier A.M., Baseman J.B. Attachment of *Mycoplasma pneumoniae* to respiratory epithelium // Infect. Immun. – 1976. – Vol. 13. – No. 3. – P. 959–966.
  124. Preston C.M., Wu K.Y., Molinski T.F., DeLong, E. F. A psychrophilic crenarchaeon inhabits a marine sponge: *Cenarchaeum symbiosum* gen. nov., sp. nov // Proceedings of the National Academy of Sciences. – 1996. – Vol. 93. – No. 13. – P. 6241–6246.
  125. Prokofeva M. I., Kostrikina N.A., Kolganova T.V., Tourova T.P., Lysenko A.M., Lebedinsky A.V., Bonch-Osmolovskaya E.A. Isolation of the anaerobic thermoacidophilic crenarchaeote *Acidilobus saccharovorans* sp. nov. and proposal of *Acidilobales* ord. nov., including *Acidilobaceae* fam. nov. and *Caldisphaeraceae* fam. nov // International journal of systematic and evolutionary microbiology. – 2009. – Vol. 59. – No. 12. – P. 3116–3122.
  126. Puigbò P., Guzmán E., Romeu A., Garcia-Vallve S. OPTIMIZER: a web server for optimizing the codon usage of DNA sequences // Nucleic acids research. – 2007. – Vol. 35. – No. suppl. 2. – P. W126–W131.
  127. Qu X., Wen J.D., Lancaster L., Noller H.F., Bustamante C., Tinoco I. The ribosome uses two active mechanisms to unwind messenger RNA during translation // Nature. – 2011. – Vol. 475. – No. 7354. – P. 118–121.
  128. Ran W., Higgs P.G. Contributions of speed and accuracy to translational selection in bacteria // PLoS One. – 2012. – Vol. 7. – No. 12. – P. e51652.
  129. dos Reis M., Wernisch L., Savva R. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome // Nucleic acids research. – 2003. – Vol. 31. – No. 23. – P. 6976–6985.
  130. dos Reis M., Savva R., Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection // Nucleic acids research. – 2004. – Vol. 32. – No. 17. – P. 5036–5044.

131. Reuveni S., Meilijson I., Kupiec M., Ruppin E., Tuller, T. Genome-scale analysis of translation elongation with a ribosome flow model // PLoS computational biology. – 2011. – Vol. 7. – No. 9. – P. e1002127.
132. Richardson S.M. Wheelan S.J., Yarrington R.M., Boeke J.D. GeneDesign: rapid, automated design of multikilobase synthetic genes // Genome research. – 2006. – Vol. 16. – No. 4. – P. 550–556.
133. Rocha E.P.C., Danchin A., Viari A. Translation in *Bacillus subtilis*: roles and trends of initiation and termination, insights from a genome analysis // Nucleic acids research. – 1999. – Vol. 27. – No. 17. – P. 3567–3576.
134. Roh S.W., Nam Y.D., Chang H.W., Sung Y., Kim K.H., Oh H.M., Bae J.W. *Halalkalicoccus jeotgali* sp. nov., a halophilic archaeon from shrimp jeotgal, a traditional Korean fermented seafood // International journal of systematic and evolutionary microbiology. – 2007. – Vol. 57. – No. 10. – P. 2296–2298.
135. Romero H., Zavala A., Musto H. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces // Nucleic Acids Research – 2000. – Vol. 28. – No. 10. – P. 2084–2090.
136. Sako Y., Nomura N., Uchida A., Ishida Y., Morii H., Koga Y., Hoaki T., Maruyama T. *Aeropyrum pernix* gen. nov., sp. nov., a novel aerobic hyperthermophilic archaeon growing at temperatures up to 100°C // International journal of systematic bacteriology. – 1996. – Vol. 46. – No. 4. – P. 1070–1077.
137. Sako Y., Nunoura T., Uchida A. *Pyrobaculum oguniense* sp. nov., a novel facultatively aerobic and hyperthermophilic archaeon growing at up to 97 degrees C // International journal of systematic and evolutionary microbiology. – 2001. – Vol. 51. – No. 2. – P. 303–309.
138. Salgado H., Moreno-Hagelsieb G., Smith T., Collado-Vides J. Operons in *Escherichia coli*: Genomic analyses and predictions // PNAS. – 2000. – Vol. 97. – No. 12. – P. 6652–6657.
139. Sankoff D., Kruskal J.B., Mainville S., Cedergren R.J. Fast algorithms to determine RNA secondary structures containing multiple loops // Time warps, string edits, and macromolecules: the theory and practice of sequence comparison/edited by David Sankoff and Joseph B. Krustal. – 1983.
140. Schleper C., Pühler G., Klenk H.P., Zillig W. *Picrophilus oshimae* and *Picrophilus torridus* fam. nov., gen. nov., sp. nov., two species of hyperacidophilic, thermophilic, heterotrophic, aerobic archaea // International journal of systematic bacteriology. – 1996. – Vol. 46. – No. 3. – P. 814–816.

141. Schmittgen T.D., Danenberg K.D., Horikoshi T., Lenz H.J., Danenberg P.V. Effect of 5-fluoro- and 5-bromouracil substitution on the translation of human thymidylate synthase mRNA // *Journal of Biological Chemistry*. – 1994. – Vol. 269. – No. 23. – P. 16269–16275.
142. Schouten S., Baas M., Hopmans E.C., Reysenbach A.L., Damsté J.S.S. Tetraether membrane lipids of *Candidatus "Aciduliprofundum boonei"*, a cultivated obligate thermoacidophilic euryarchaeote from deep-sea hydrothermal vents // *Extremophiles*. – 2008. – Vol. 12. – No. 1. – P. 119–124.
143. Seffens W., Digby D. mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences // *Nucleic acids research*. – 1999. – Vol. 27. – No. 7. – P. 1578–1584.
144. Seob B.S., Kim Y.J., Yang S.H., Lim J.K., Jeon J.H., Lee H.S., Kang, S.G., Kim, S.J., Lee J.H. *Thermococcus onnurineus* sp. nov., a hyperthermophilic archaeon isolated from a deep-sea hydrothermal vent area at the PACMANUS field // *Journal of microbiology and biotechnology*. – 2006. – Vol. 16. – No. 11. – P. 1826–1831.
145. Shah P., Ding, Y., Niemczyk M., Kudla G., Plotkin, J.B. Rate-limiting steps in yeast protein translation // *Cell*. – 2013. – Vol. 153. – No. 7. – P. 1589–1601.
146. Sharma A.K., Chowdhury D. Distribution of dwell times of a ribosome: effects of infidelity, kinetic proofreading and ribosome crowding // *Physical biology*. – 2011. – Vol. 8. – No. 2. – P. 026005.
147. Sharp, P.M., Tuohy T.M., Mosurski K.R. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes // *Nucleic acids research*. – 1986. – Vol. 14. – No. 13. – P. 5125–5143.
148. Sharp P.M., Li W.H. An evolutionary perspective on synonymous codon usage in unicellular organisms // *Journal of molecular evolution*. – 1986. – Vol. 24. – No. 1–2. – P. 28–38.
149. Sharp P.M., Li W.H. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications // *Nucleic Acids Res*. – 1987. – Vol. 15. – P. 1281–1295.
150. Siller E., DeZwaan D.C., Anderson J.F., Freeman B.C., Barral J.M. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency // *Journal of molecular biology*. – 2010. – Vol. 396. – No. 5. – P. 1310–1318.
151. de Smit M.H., van Duin J. Control of translation by mRNA secondary structure in *Escherichia coli*: A quantitative analysis of literature data // *Journal of molecular biology*. – 1994. – Vol. 244. – No. 2. – P. 144–150.
152. Sorensen M.A., Kurland C.G., Pedersen S // Codon usage determines translation rate in *Escherichia coli*. // *J. Mol. Biol*. – 1989. – Vol. 207. – P. 365–377.

153. Steinsbu B.O., Thorseth I.H., Nakagawa S., Inagaki F., Lever M.A., Engelen B., Øvreås L., Pedersen R.B. *Archaeoglobus sulfaticallidus* sp. nov., a thermophilic and facultatively lithoautotrophic sulfate-reducer isolated from black rust exposed to hot ridge flank crustal fluids // *International journal of systematic and evolutionary microbiology*. – 2010. – Vol. 60. – No. 12. – P. 2745–2752.
154. Stenico M., Lloyd A.T., Sharp P.M. Codon usage in *Caenorhabditis elegans*: delineation of translational selection and mutational biases // *Nucleic Acids Res.* – 1994. – Vol. 22. – P. 2437–2446.
155. Studer S.M., Joseph S. Unfolding of mRNA secondary structure by the bacterial translation initiation complex // *Molecular cell*. – 2006. – Vol. 22. – No. 1. – P. 105–115.
156. Suzuki T., Iwasaki T., Uzawa T., Hara K., Nemoto N., Kon T., Ueki T., Yamagishi A., Oshima T. *Sulfolobus tokodaii* sp. nov. (f. *Sulfolobus* sp. strain 7), a new member of the genus *Sulfolobus* isolated from Beppu Hot Springs, Japan // *Extremophiles*. – 2002. – Vol. 6. – No. 1. – P. 39–44.
157. Swainston N., Currin A., Day P.J., Kell D.B. GeneGenie: optimized oligomer design for directed evolution // *Nucleic acids research*. – 2014. – P. gku336.
158. Takyar S., Hickerson R.P., Noller H.F. mRNA helicase activity of the ribosome // *Cell*. – 2005. – Vol. 120. – No. 1. – P. 49–58.
159. Thanaraj T.A., Argos P. Ribosome-mediated translational pause and protein domain organization // *Protein Science*. – 1996. – Vol. 5. – No. 8. – P. 1594–1612.
160. Tinoco I., Uhlenbeck O.C., Levine M.D. Estimation of secondary structure in ribonucleic acids // *Nature*. – 1971. – Vol. 230. – No. 5293. – P. 362–367.
161. Tinoco I., Borer P.N., Dengler B., Levine M.D., Uhlenbeck O.C., Crothers D.M., Gralla, J. Improved estimation of secondary structure in ribonucleic acids // *Nature*. – 1973. – Vol. 246. – No. 150. – P. 40–41.
162. Torreblanca M., Rodriguez-Valera F., Juez G., Ventosa A., Kamekura M., Kates M. Classification of non-alkaliphilic halobacteria based on numerical taxonomy and polar lipid composition, and description of *Haloarcula* gen. nov. and *Haloferax* gen. nov // *Systematic and Applied Microbiology*. – 1986. – Vol. 8. – No. 1. – P. 89–99.
163. Tuller T., Carmi A., Vestsigian K., Navon S., Dorfan Y., Zaborske J., Pan T., Dahan O., Furman I., Pilpel Y. An evolutionarily conserved mechanism for controlling the efficiency of protein translation // *Cell*. – 2010. – Vol. 141. – No. 2. – P. 344–354.
164. Tuller T., Veksler-Lublinsky I., Gazit N., Kupiec M., Ruppin E., Ziv-Ukelson M. Composite effects of gene determinants on the translation speed and density of ribosomes // *Genome Biol.* – 2011. – Vol. 12. – No. 11. – P. R110.



165. Uenoyama A., Miyata M. Identification of a 123-kilodalton protein (Gli123) involved in machinery for gliding motility of *Mycoplasma mobile* // Journal of bacteriology. – 2005. – Vol. 187. – No. 16. – P. 5578–5584.
166. Uhlenbeck O.C., Borer P.N., Dengler B., Tinoco I. Stability of RNA hairpin loops: A 6-Cm-U 6 // Journal of molecular biology. – 1973. – Vol. 73. – No. 4. – P. 483–496.
167. Varenne S., Buc J., Lloubes R., Lazdunski C. Translation is a non-uniform process. Effect of tRNA availability on the rate of elongation of nascent polypeptide chains // J. Mol. Biol. – 1984. – Vol. 180. – P. 549–576.
168. Villalobos A., Ness J.E., Gustafsson C., Minshull J., Govindarajan S. Gene Designer: a synthetic biology tool for constructing artificial DNA segments // BMC bioinformatics. – 2006. – Vol. 7. – No. 1. – P. 285.
169. Vladimirov N.V., Likhoshvai V.A., Matushkin Yu.G. Correlation of codon biases and potential secondary structures with mRNA translation efficiency in unicellular organisms // Mol. Biol. – 2007. – Vol. 41. – No. 5. – P. 926–933.
170. Vogel C., Abreu R. de S., Ko D., Le S.-Y., Shapiro B.A., Burns S.C., Sandhu D., Boutz D.R., Marcotte E.M., Penalva L.O. Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line // Molecular systems biology. – 2010. – Vol. 6. – No. 1.
171. Vökl P., Huber R., Drobner E., Rachel R., Burggraf S., Trincone A., Stetter, K.O. *Pyrobaculum aerophilum* sp. nov., a novel nitrate-reducing hyperthermophilic archaeum // Applied and Environmental Microbiology. – 1993. – Vol. 59. – No. 9. – P. 2918–2926.
172. Wada K. S., Aota R., Tsuchiya F., Ishibashi T., Gojobori T., Ikemura T. Codon usage tabulated from GenBank genetic sequence data // Nucleic Acids Res. – 1990. – Vol. 18. suppl. – P. 2367–2411.
173. Wainø M., Tindall B. J., Ingvorsen K. *Halorhabdus utahensis* gen. nov., sp. nov., an aerobic, extremely halophilic member of the Archaea from Great Salt Lake, Utah // International Journal of Systematic and Evolutionary Microbiology. – 2000. – Vol. 50. – No. 1. – P. 183–190.
174. Waites K.B., Balish M.F., Atkinson T.P. New insights into the pathogenesis and detection of *Mycoplasma pneumoniae* infections // Future Microbiol. – 2008. – Vol. 3. – No. 6. – P. 635–648.
175. Waldman Y.Y., Tuller T., Shlomi T., Sharan R., Ruppin E. Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages // Nucleic acids research. – 2010. – Vol. 38. – No. 9. – P. 2964–2974.

176. Waldo R. H., Krause D. C. Synthesis, stability, and function of cytoadhesin P1 and accessory protein B/C complex of *Mycoplasma pneumoniae* // Journal of bacteriology. – 2006. – Vol. 188. – No. 2. – P. 569–575.
177. Waterman M.S. Secondary structure of single-stranded nucleic acids // Adv. math. suppl. studies. – 1978. – Vol. 1. – P. 167–212.
178. Waterman M.S., Smith T.F. RNA secondary structure: A complete mathematical analysis // Mathematical Biosciences. – 1978. – Vol. 42. – No. 3. – P. 257–266.
179. (A) Welch M., Villalobos A., Gustafsson C., Minshull J. You're one in a googol: optimizing genes for protein expression // Journal of the Royal Society Interface. – 2009. – P. rsif.2008.0520. focus.
180. (B) Welch M., Govindarajan S., Ness J.E., Villalobos A., Gurney A., Minshull J., Gustafsson C. Design parameters to control synthetic gene expression in *Escherichia coli* // PloS one. – 2009. – Vol. 4. – No. 9. – P. e7002.
181. Wen J.D., Lancaster L., Hodges C., Zeri A.C., Yoshimura S.H., Noller H.F., Bustamante C., Tinoco I. Following translation by single ribosomes one codon at a time // Nature. – 2008. – Vol. 452. – No. 7187. – P. 598–603.
182. Workman C., Krogh A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution // Nucleic Acids Research. – 1999. – Vol. 27. – No. 24. – P. 4816–4822.
183. Wright F., The 'effective number of codons' used in a gene // Gene. – 1990. – Vol. 87. – No. 1. – P. 23–29.
184. Xia X., MacKay V., Yao X., Wu J., Miura F., Ito T., Morris D. R. Translation initiation: a regulatory role for poly (A) tracts in front of the AUG codon in *Saccharomyces cerevisiae* // Genetics. – 2011. – Vol. 189. – No. 2. – P. 469–478.
185. Xie P. Model of ribosome translation and mRNA unwinding // European Biophysics Journal. – 2013. – Vol. 42. – No. 5. – P. 347–354.
186. Yang J.R., Chen X., Zhang J. Codon-by-codon modulation of translational speed and accuracy via mRNA folding // PLoS biology. – 2014. – Vol. 12. – No. 7. – P. e1001910.
187. Yanofsky C., van Cleemput M. Nucleotide sequence of *trpE* of *Salmonella typhimurium* and its homology with the corresponding sequence of *Escherichia coli* // Journal of molecular biology. – 1982. – 155(3), 235–246.
188. Yasuda M., Oyaizu H., Yamagishi A., Oshima T. Morphological variation of new *Thermoplasma acidophilum* isolates from Japanese hot springs // Applied and environmental microbiology. – 1995. – Vol. 61. – No. 9. – P. 3482–3485.

189. You X.Y., Liu C., Wang S.Y., Jiang C.Y., Shah S.A., Prangishvili D., She Q., Liu S.-J., Garrett R.A. Genomic analysis of *Acidianus hospitalis* W1 a host for studying crenarchaeal virus and plasmid life cycles // *Extremophiles*. – 2011. – Vol. 15. – No. 4. – P. 487–497.
190. Zama M. Correlation between mRNA structure of the coding region and translational pauses // *Nucleic acids symposium series*. – Oxford University Press. – 1999. – Vol. 42. – No. 1. – P. 81–82.
191. Zhang G., Ignatova Z. Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis // *PLoS One*. – 2009. – Vol. 4. – No. 4. – P. e5036.
192. Zhang G., Hubalewska M., Ignatova Z. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding // *Nature structural & molecular biology*. – 2009. – Vol. 16. – No. 3. – P. 274–280.
193. Zheng Q., Ryvkin P., Li F., Dragomir I., Valladares O., Yang J., Cao K., Wang L.-S., Gregory B.D. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis* // *PLoS genetics*. – 2010. – Vol. 6. – No. 9. – P. e1001141.
194. Zillig W., Holz I., Wunderl S. NOTES: *Hyperthermus butylicus* gen. nov., sp. nov., a Hyperthermophilic, Anaerobic, Peptide-Fermenting, Facultatively H<sub>2</sub>S-Generating Archaeobacterium // *International journal of systematic bacteriology*. – 1991. – Vol. 41. – No. 1. – P. 169–170.
195. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction // *Nucleic acids research*. – 2003. – Vol. 31. – No. 13. – P. 3406–3415.
196. Zuker M., Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information // *Nucleic acids research*. – 1981. – Vol. 9. – No. 1. – P. 133–148.

ПРИЛОЖЕНИЕ

Таблица 1. Значения типов индекса EEI для 2582 организмов из домена Бактерии.

M1-M5 – значения средних положений (рангов) генов рибосомных белков в отсортированных по увеличению индекса EEI списках генов организмов, R1-R5 – стандартные отклонения от средних (см. главу «Методы и алгоритмы»). Синим цветом выделены наибольшие значения параметра M для организма, красным – организмы с максимальным параметром M < 30.

Table with columns: Organism, Index, M1, R1, M2, R2, M3, R3, M4, R4, M5, R5. The table lists 2582 bacterial organisms and their corresponding EEI index values and standard deviations across five different gene sets (M1-M5, R1-R5).



Таблица 1. (Продолжение).

Organism	Index	M1-M10										Organism	Index	M11-M20									
		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10			M11	M12	M13	M14	M15	M16	M17	M18	M19	M20
<i>Bacillus amyloliquefaciens</i> subsp. <i>plantarum</i> ucmbr	1	86	36	25	61	27	44	50	57	75	83	30	50	47	-31	62	82	32	-19	63			
<i>Corynebacterium glutamicum</i> n8001	1	86	36	-29	58	-32	54	65	44	40	83	30	49	47	-31	62	82	32	-19	63			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>agona</i>	1	86	36	25	61	27	44	50	57	75	83	31	-14	51	-13	52	67	34	58	44			
<i>Streptococcus suis</i> s375	1	86	36	-17	64	-18	59	74	41	50	83	31	49	48	-30	64	78	33	-20	62			
<i>Escherichia coli</i> dh1	1	86	37	12	66	15	54	81	42	78	83	31	50	46	-33	59	78	34	-16	66			
<i>Escherichia coli</i> p12b	1	86	37	11	66	13	54	66	52	62	83	32	10	66	36	56	73	45	72	28			
<i>Escherichia coli</i> um146	1	86	37	7	64	7	54	77	42	70	83	32	34	56	-47	63	76	34	-44	59			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>pullorum</i>	1	86	37	22	62	23	45	79	42	75	83	32	46	46	-30	62	78	34	-15	62			
<i>Streptococcus mitis</i> b6	1	86	37	-3	55	-4	58	80	38	16	83	33	-6	61	-11	63	82	28	53	56			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>bareilly</i>	1	86	38	22	64	28	46	80	43	77	83	34	-7	63	-12	64	82	28	53	56			
<i>Streptococcus anginosus</i> s238	1	86	38	24	60	-30	59	69	46	-26	83	35	-8	67	-8	64	75	44	52	62			
<i>Vibrio campbellii</i> atcc hsa-1116	1	86	38	-27	61	-33	58	66	46	39	83	35	6	64	17	63	67	50	28	47			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>heidelberg</i>	1	86	40	21	65	27	46	81	47	83	83	35	39	44	-47	55	41	48	-56	52			
<i>Streptococcus pyogenes</i> subsp. <i>pyogenes</i> ecm3	1	86	40	-5	61	-21	66	77	46	22	83	35	19	57	-49	52	76	37	-37	63			
<i>Streptococcus pneumoniae</i> mv200	1	86	40	-2	58	-9	56	74	43	-3	83	35	17	57	11	56	43	58	10	51			
<i>Streptococcus suis</i> s9	1	86	41	-17	62	-28	59	81	43	60	83	36	-18	61	45	42	50	47	81	34			
<i>Pseudomonas fluorescens</i> Pf0-1	1	85	17	25	65	25	57	80	24	77	83	36	39	53	-42	46	79	37	82	35			
<i>Lactobacillus rhamnosus</i> lock900	1	85	18	-5	54	-18	55	41	46	32	83	37	9	61	13	55	23	57	18	51			
<i>Lactobacillus rhamnosus</i> lock908	1	85	18	-6	53	-20	56	47	43	40	83	37	-8	60	-1	58	73	44	62	54			
<i>Acetobacter pasteurianus</i> 3866	1	85	19	23	58	-7	56	82	30	66	83	38	16	60	39	39	-6	57	44	35			
<i>Hydrobacter polytropus</i> dm 29226	1	85	19	24	59	-22	65	78	20	-13	83	38	19	62	-24	58	69	46	-20	57			
<i>Polynucleobacter necessarius</i> subsp. <i>necessarius</i> s1	1	85	19	3	59	-1	62	70	39	47	83	39	20	55	-21	61	78	40	-12	63			
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> s1272	1	85	22	47	47	-32	60	79	31	-23	83	39	11	63	11	53	79	41	76	42			
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> s1272	1	85	23	50	46	-32	62	76	32	-27	83	39	-30	58	-1	52	10	62	41	69			
<i>Baumannella thailandica</i> 5031	1	85	24	21	56	-39	57	81	33	-12	83	40	13	61	42	40	-15	53	48	35			
<i>Staphylococcus aureus</i> subsp. <i>aureus</i> us300 fpr	1	85	24	24	60	47	-32	63	31	63	83	40	22	60	47	-32	63	31	63	31			
<i>Rhodospirillum rubrum</i> ssp. <i>rubrum</i> nml	1	85	25	24	56	37	45	77	28	81	83	42	20	59	32	42	-2	53	43	38			
<i>Mycobacterium abscessus</i> subsp. <i>abscessus</i> h39594	1	85	27	7	56	4	54	81	34	62	83	42	22	55	-16	61	71	43	-12	62			
<i>Bacteroides xylophilus</i> sbl1	1	85	28	-7	56	2	57	76	35	65	83	43	12	64	13	54	76	46	70	48			
<i>Enterobacter cloacae</i> subsp. <i>disseverans</i> sdm	1	85	28	22	58	47	39	67	42	83	83	45	16	66	11	56	79	45	68	54			
<i>Baumannella thailandica</i> 50310	1	85	28	24	53	-31	53	81	31	10	83	46	-7	69	-9	59	81	44	5	53			
<i>Staphylococcus aureus</i> sfn9393	1	85	28	47	47	-29	63	80	28	-12	82	16	7	56	7	56	77	29	65	39			
<i>Zymomonas mobilis</i> subsp. <i>mobilis</i> atcc 10988	1	85	28	16	59	-33	59	63	45	-13	82	17	11	57	10	53	79	25	71	35			
<i>Arthro bacter aureus</i> s1	1	85	29	25	55	-31	54	75	37	71	82	22	16	59	-3	62	73	32	38	55			
<i>Arthrobacter parvulus</i> dm 20469	1	85	29	2	62	-50	55	59	44	-48	82	22	0	62	-12	58	33	61	25	66			
<i>Bacillus melanosus</i> n1	1	85	29	11	57	14	60	75	38	63	82	26	0	58	-33	56	72	34	24	58			
<i>Serratia proteamaculans</i> 568	1	85	29	43	51	58	42	68	39	85	82	24	11	53	3	52	77	28	53	49			
<i>Enterobacter asburiae</i> ifa	1	85	30	17	67	43	46	57	55	78	82	25	36	57	47	40	64	39	82	25			
<i>Roseburia hominis</i> a2-183	1	85	32	20	51	47	42	15	52	20	82	25	46	48	-33	62	80	29	-19	63			
<i>Streptococcus sp.</i> s2	1	85	32	-12	61	12	63	39	58	20	82	25	46	48	-33	62	80	29	-19	63			
<i>Streptococcus oligofermentans</i> as 1_3089	1	85	35	-6	54	-3	53	78	39	51	82	26	0	58	-33	56	72	34	24	58			
<i>Bacillus infantis</i> nml b-14911	1	85	36	25	63	31	51	21	60	19	82	26	10	51	-1	53	73	35	49	51			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>typhimurium</i>	1	85	36	22	63	28	46	80	44	81	82	26	9	51	-1	53	72	36	49	51			
<i>Escherichia coli</i> o157:h7 str 12009	1	85	37	12	63	12	52	80	40	78	82	26	47	48	-33	63	80	30	-19	64			
<i>Streptococcus gallolyticus</i> snc34	1	85	37	14	59	-37	63	82	39	19	82	29	48	46	-30	63	79	30	-18	61			
<i>Streptococcus pyogenes</i> ssp. <i>pyogenes</i> maffedoi	1	85	37	-1	65	-22	62	78	39	81	82	30	55	45	-34	58	82	37	-28	57			
<i>Vibrio anguillarum</i> 775	1	85	37	-10	59	-15	62	76	43	42	82	31	35	47	-10	59	81	32	4	59			
<i>Escherichia coli</i> k011f	1	85	38	10	65	10	53	79	44	76	82	31	-15	61	15	63	21	62	44	64			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>heidelberg</i>	1	85	40	18	63	24	45	76	48	76	82	31	27	60	-12	62	67	46	6	72			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>heidelberg</i>	1	85	40	18	63	24	45	76	48	76	82	31	27	60	-12	62	67	46	6	72			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>typhimurium</i>	1	85	40	19	64	26	47	78	47	75	82	31	32	56	-23	65	78	35	17	70			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>typhimurium</i>	1	85	40	20	64	26	46	78	47	77	82	32	46	47	-32	62	75	38	-17	66			
<i>Streptococcus thermophilus</i> nd03	1	85	40	-10	53	-37	59	82	41	37	82	33	11	61	-12	56	1	52	6	65			
<i>Streptococcus pneumoniae</i> tch84119a	1	85	42	4	57	-3	55	82	38	11	82	33	21	55	-3	69	80	30	41	62			
<i>Streptococcus suis</i> hml07	1	85	42	-1	64	-3	53	78	39	51	82	34	13	63	37	41	-11	59	52	35			
<i>Salmonella enterica</i> subsp. <i>enterica</i> serovar <i>agona</i>	1	85	43	22	62	24	46	82	46	82	82	34	13	59	-25	54	71	39	44	56			
<i>Bifidobacterium longum</i> subsp. <i>infantis</i> 157f	1	84	14	9	56	8	53	80	27	68	82	35	15	53	-20	63	77	33	-2	58			
<i>Bifidobacterium longum</i> subsp. <i>longum</i> kacc_915	1	84	14	8	55	7	53	81	25	71	82	35	25	52	-27	60	65	45	-29	57			
<i>Acetobacter pasteurianus</i> ssp. 3283-01	1	84	20	21	58	9	56	79	34	64	82	35	3	65	6	58	63	48	60	52			
<i>Acetobacter pasteurianus</i> ssp. 3283-12	1	84	20	21	58	9	56	79	34	64	82	35	3	65	6	58	63	48	60	52			
<i>Acetobacter pasteurianus</i> ssp. 3283-07	1	84	20	21	58	9	56	79	34	64	82	36	-2	55	-47	54	81	35	19	69			
<i>Acetobacter pasteurianus</i> ssp. 3283-12	1	84	20	21	58	9	56	79	34	64	82	37	23	57	-20	59	80	36	-11	61			
<i>Acetobacter pasteurianus</i> ssp. 3283-22	1	84	20	21	58	9	56	79	34	64	82	37	2	65	-15	57	60	48	34	64			
<i>Acetobacter pasteurianus</i> ssp. 3283-26	1	84	20	21	58	9	56	79	34	64	82	38	0	56	-44	59	81	36	21	71			
<i>Acetobacter pasteurianus</i> ssp. 3283-12	1	84	20	21	58	9	56	79	34	64	82	39	13	63	37	41	-11	59	52	35			
<i>Acetobacter pasteurianus</i> ssp. 3283-01-42c	1	84	21	22	58	9	56	79	34	65	82	39	7	66	-17	64	71	32	5	62			
<i>Listeria monocytogenes</i> 08-923	1	84	22	31	60	-22	61	83	24	5	82	40	-10	64	-15	64	77	37	47	58			
<i>Listeria monocytogenes</i> 08-5578	1	84	22	31	60	-22	60	83	23	-13	82	40	23	62	27	47	47	57	67	46			
<i>Staphylococcus aureus</i> ml	1	84	22	51	47	-32	61	84	24	-20	82	41	21	53	-18								



Таблица 1. (Продолжение).

Organism	Index	M1										M2										M3										M4										M5																		
		M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10									
<i>mycobacterium goodii</i>	1	80	27	80	50	332	48	69	38	64	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41
<i>pseudomonas syringae_pv._plasmocohv_1448a</i>	1	80	29	33	53	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41
<i>staphylococcus aureus_subsp._aureus_eid98</i>	1	80	32	50	47	-31	62	79	31	-22	61	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>listeria monocytogenes_scc2372</i>	1	80	33	28	59	-26	62	79	33	9	60	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>listeria monocytogenes_scc2479</i>	1	80	33	28	59	-26	62	79	33	9	60	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>listeria monocytogenes_scc1719</i>	1	80	34	28	60	-27	61	79	33	9	58	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>staphylococcus aureus_subsp._aureus_lap251</i>	1	80	34	44	51	-29	63	80	37	-15	65	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>velocella parvula_dsm_2008</i>	1	80	34	23	57	-2	67	78	31	47	56	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>soltika canadensis_dsm_3403</i>	1	80	37	10	64	-32	57	74	50	-5	62	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>staphylococcus epidermidis_atcc_12228</i>	1	80	38	39	51	-23	64	79	33	-21	61	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>bacillus thuringiensis_0125</i>	1	80	39	17	62	-10	63	65	49	57	58	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>propionibacterium acnes_H969a1</i>	1	80	40	-8	63	-14	65	77	37	45	61	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>akeromonas mucloidi_atcc_27126</i>	1	80	41	-7	67	-18	64	68	48	44	67	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>akeromonas mucloidi_str_‘bakare_sea_4d45’</i>	1	80	41	-9	65	-14	62	71	47	48	64	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>bacillus arylohalophilus_subsp._plantarum_str_1</i>	1	80	41	17	61	38	40	-12	53	56	45	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>bacillus halodurans_0125</i>	1	80	41	-29	58	-25	57	-1	60	-24	59	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>akeromonas mucloidi_str_‘omian_sea_umbf’</i>	1	80	42	-8	62	-11	60	75	43	49	60	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>bacillus arylohalophilus_subsp._plantarum_nau-1</i>	1	80	42	21	60	32	41	76	43	79	39	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>bacillus cereus_nc7401</i>	1	80	42	38	44	-40	55	70	40	-36	56	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>acetobacter baumannii_sdf</i>	1	80	43	5	68	-35	59	75	49	35	66	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>akeromonas mucloidi_dh1e1</i>	1	80	43	-8	63	-9	60	76	43	50	67	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>akeromonas mucloidi_str_‘omian_sea_uf’</i>	1	80	43	-10	61	-13	59	73	45	48	64	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>streptococcus pneumoniae_at026</i>	1	80	46	-6	61	-4	60	75	46	13	57	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>haemophilus parasvis_s10165</i>	1	80	48	-7	58	-33	55	73	49	44	65	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>bifidobacterium longum_nc2705</i>	1	79	24	12	58	-13	54	75	32	65	62	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>mycobacterium goodii_str_‘omian_sea_umbf’</i>	1	79	26	32	48	-28	48	69	36	59	47	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>chlamydia trachomatis</i>	1	79	27	-8	65	-12	52	68	41	13	57	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>chelatovirus_sp_bcl1</i>	1	79	28	-17	61	-15	63	45	54	29	67	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>rhotzobium leguminosarum_bv_trifolii_wsm2304</i>	1	79	29	24	59	34	51	73	34	78	29	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>strobilobium fredii_usda_257</i>	1	79	31	3	63	16	58	20	65	34	62	41	38	31	63	46	47	63	44	69	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41	41	38	31	63	46	47	63	44	69	41										
<i>agrobacterium radiobacter_384</i>	1	79	33	30	53	-14	49	42	49	75	38	41	38	31	63	46	47	63	44	69																																								

Таблица 1. (Продолжение).

Table with columns: Organism, Index, M1, R1, M2, R2, M3, R3, M4, R4, M5, R5, Organism, Index, M1, R1, M2, R2, M3, R3, M4, R4, M5, R5. The table lists various bacterial species and their corresponding index values across multiple columns.



Таблица 1. (Продолжение).

Table with columns: Organism, Index, M1, R1, M2, R2, M3, R3, M4, R4, M5, R5, Organism, Index, M1, R1, M2, R2, M3, R3, M4, R4, M5, R5. The table lists various bacterial species and their corresponding index values across multiple columns.









Таблица 1. (Продолжение).

Organism	Index	Organism															
		M1	R1	M2	R2	M3	R3	M4	R4	M5	R5	M6	R6	M7	R7		
<i>Actinobacterium</i> sp. <i>h101</i>	4	6	61	29	55	38	60	38	47	52	34	59	30	49	30	48	25
<i>Candidata</i> <i>blachmannii</i> <i>flouidans</i>	4	16	51	28	58	4	68	38	48	7	61	59	30	49	30	78	24
<i>Desulfovibrio africanus</i> <i>str. walis bay</i>	4	-7	62	37	46	30	51	38	52	36	56	47	60	63	33	68	27
<i>Calovirus</i> <i>sp. pcc_7507</i>	4	-29	49	29	56	-27	64	38	54	-39	51	59	30	49	30	85	32
<i>Dichelobacter nodosus</i> <i>ves1705a</i>	4	-2	62	33	57	30	50	38	55	35	47	63	34	51	48	65	34
<i>Mycobacterium tuberculosis</i> <i>rgb27</i>	4	-6	58	16	55	3	59	38	61	26	64	59	30	49	35	81	40
<i>Mycobacterium tuberculosis</i> <i>rgb423</i>	4	-13	57	14	55	3	58	38	61	28	67	63	30	49	30	81	26
<i>Mycobacterium tuberculosis</i> <i>h37ra</i>	4	-2	60	16	55	5	61	38	62	28	70	60	30	49	41	68	28
<i>Mycobacterium tuberculosis</i> <i>str. beijingn1203</i>	4	-10	60	18	55	6	61	38	63	28	68	63	30	49	48	60	37
<i>Mycobacterium bovis</i> <i>bgc_str. pasteur_1175p2</i>	4	0	61	16	55	7	62	38	64	27	68	63	30	49	46	48	40
<i>Acetobacterium arabaticum</i> <i>dm_5501</i>	4	18	61	35	52	-9	64	37	53	-10	65	60	30	49	42	65	33
<i>Thermotoga</i> <i>sp. x514</i>	4	-5	58	35	54	-5	64	37	59	-11	59	60	30	49	42	65	33
<i>Desulfohalobium</i> <i>merdii</i> <i>dm_13257</i>	4	-28	56	31	56	-20	64	37	59	-35	58	62	30	49	46	39	81
<i>Desulfohalobium</i> <i>autotrophicum</i> <i>hm2</i>	4	17	59	35	60	25	56	37	60	23	51	60	30	49	42	65	33
<i>Blattabacterium</i> <i>sp. (blattella germanica) str. bge</i>	4	12	62	33	54	13	65	37	61	15	61	63	30	49	45	55	36
<i>Mycobacterium tuberculosis</i> <i>cds5079</i>	4	22	60	15	55	34	61	37	62	25	69	60	30	49	42	67	41
<i>Serratia symbiotica</i> <i>str. 'causa cedii'</i>	4	22	68	26	60	-20	68	37	63	-10	74	60	30	49	47	67	38
<i>Mycobacterium tuberculosis</i> <i>h37rv</i>	4	-5	59	18	56	8	61	37	63	31	69	60	30	49	45	66	31
<i>Helicobacter</i> <i>chnaidi</i> <i>atcc_baa-847</i>	4	-16	56	34	48	2	58	36	51	-11	55	60	30	49	44	29	80
<i>Mycoplasma parvum</i> <i>str. andana</i>	4	-8	52	35	38	6	57	36	53	-3	52	60	30	49	41	71	29
<i>Candidata</i> <i>solibacter</i> <i>solitans</i> <i>dm6076</i>	4	22	51	26	53	17	60	36	54	33	61	60	30	49	42	65	33
<i>Halobacterium</i> <i>halobium</i> <i>sll</i>	4	12	57	18	55	18	55	36	54	32	57	60	30	49	42	65	33
<i>Pebobacter</i> <i>carbinolicus</i> <i>dm_2380</i>	4	25	43	25	57	21	55	36	58	36	58	60	30	49	42	65	33
<i>Nostoc</i> <i>sp. pcc_7107</i>	4	-29	58	32	61	-43	61	36	60	-48	54	60	30	49	45	63	35
<i>Lasiosira atricellaris</i> <i>rh43</i>	4	20	66	30	63	14	61	36	61	10	57	60	30	49	45	63	35
<i>Desulfohalobium</i> <i>acetoxidans</i> <i>dm_771</i>	4	18	61	35	52	-9	64	37	53	-10	65	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>cdsnt206</i>	4	11	59	15	55	3	61	36	62	26	70	60	30	49	42	65	33
<i>Blattabacterium</i> <i>sp. (blattella germanica)</i>	4	-11	64	31	53	34	61	36	63	23	64	60	30	49	40	62	35
<i>Mycobacterium bovis</i> <i>bgc_str. tokyo_172</i>	4	-5	60	17	55	7	62	36	63	24	69	60	30	49	42	65	33
<i>Mycobacterium</i> <i>caenieta</i> <i>cp1_14001059</i>	4	-3	63	19	57	9	62	36	64	29	70	60	30	49	42	65	33
<i>Synechococcus</i> <i>sp. pcc_7502</i>	4	33	67	27	66	14	62	34	66	30	69	60	30	49	42	65	33
<i>Mycobacterium</i> <i>ayum</i> <i>104</i>	4	33	58	13	59	8	62	35	60	23	66	60	30	49	42	65	33
<i>Mycobacterium</i> <i>caenieta</i> <i>cp1_140070017</i>	4	-8	58	19	56	8	61	35	60	21	66	60	30	49	42	65	33
<i>Desulfohalobium</i> <i>gibsoniae</i> <i>dm_7213</i>	4	30	60	28	61	-11	56	35	63	-22	45	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>km_1435</i>	4	-4	60	16	55	4	61	35	63	25	69	60	30	49	42	65	33
<i>Wohlschia</i> <i>endosymbiont</i> <i>of</i> <i>dirosiphia</i> <i>melanos</i>	4	29	56	10	61	17	60	34	66	30	69	60	30	49	42	65	33
<i>Candidata</i> <i>profida</i> <i>armata</i>	4	-27	54	33	47	-1	64	34	50	-13	60	60	30	49	42	65	33
<i>Helicobacter</i> <i>chnaidi</i> <i>paq611</i>	4	-9	59	28	54	0	59	34	51	-6	54	60	30	49	42	65	33
<i>Candidata</i> <i>desulfonitro</i> <i>autotrophicum</i> <i>mp104c</i>	4	0	55	25	52	19	57	34	56	23	63	60	30	49	42	65	33
<i>Mycoplasma pneumoniae</i> <i>309</i>	4	-6	55	23	56	-26	56	34	60	-27	49	60	30	49	42	65	33
<i>Mesospira</i> <i>prata</i> <i>meq1_ag_4_2</i>	4	33	67	27	66	14	62	34	66	30	69	60	30	49	42	65	33
<i>Wohlschia</i> <i>sp. wri</i>	4	26	59	9	60	17	58	33	47	30	54	60	30	49	42	65	33
<i>Wohlschia</i> <i>endosymbiont</i> <i>of</i> <i>dirosiphia</i> <i>simulans_v</i>	4	21	57	11	58	12	63	33	47	16	57	60	30	49	42	65	33
<i>Wohlschia</i> <i>endosymbiont</i> <i>of</i> <i>dirosiphia</i> <i>simulans_v</i>	4	25	61	12	63	16	61	33	49	25	58	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>h37rv</i>	4	-4	60	16	55	5	61	35	63	25	69	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>7109-09</i>	4	-6	60	16	56	4	61	35	64	20	69	60	30	49	42	65	33
<i>Mycoplasma pneumoniae</i> <i>ml29</i>	4	-7	57	24	56	-26	56	32	61	-28	51	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>km_605</i>	4	-4	60	16	55	4	61	32	64	26	70	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>fl1</i>	4	-9	60	17	55	7	61	32	64	20	69	60	30	49	42	65	33
<i>Wohlschia</i> <i>endosymbiont</i> <i>of</i> <i>Brugia</i> <i>malayi</i>	4	17	61	13	63	11	56	34	69	13	64	60	30	49	42	65	33
<i>Neosporocarpus</i> <i>thermophilus</i> <i>pcc_7203</i>	4	10	55	17	59	-21	61	31	57	-25	60	60	30	49	42	65	33
<i>Nitrosospora</i> <i>multiformis</i> <i>atcc_25196</i>	4	11	61	18	69	19	60	31	58	25	59	60	30	49	42	65	33
<i>Waddlia chondrophila</i> <i>wsu_86-1044</i>	4	9	61	20	59	-6	59	31	60	-4	65	60	30	49	42	65	33
<i>Mycoplasma pneumoniae</i> <i>ml29-67</i>	4	-5	58	25	56	-27	58	31	62	-27	53	60	30	49	42	65	33
<i>Wohlschia</i> <i>endosymbiont</i> <i>of</i> <i>Chocoma</i> <i>ocherugi</i>	4	29	62	27	66	14	62	34	69	15	65	60	30	49	42	65	33
<i>Mycobacterium tuberculosis</i> <i>km_4207</i>	4	-5	61	16	55	4	62	30	64	25	70	60	30	49	42	65	33
<i>Mycoplasma pneumoniae</i> <i>fl</i>	4	-11	60	24	60	-32	57	30	67	-33	53	60	30	49	42	65	33
<i>Ehrlichia</i> <i>maris</i> <i>as145</i>	4	14	59	26	61	2	60	29	54	10	59	60	30	49	42	65	33
<i>Thermotoga</i> <i>potera</i> <i>jr</i>	4	19	61	24	60	18	61	29	58	18	60	60	30	49	42	65	33
<i>Candidata</i> <i>methylobacterium</i> <i>oxyfera</i>	4	20	60	16	52	17	49	28	58	7	60	60	30	49	42	65	33
<i>Helicobacter</i> <i>hahmannii</i> <i>asb1_4</i>	4	20	60	16	52	17	49	28	58	7	60	60	30	49	42	65	33
<i>Ehrlichia</i> <i>canis</i> <i>str. jakc</i>	4	14	63	21	60	12	61	28	59	12	66	60	30	49	42	65	33
<i>Thyocyba</i> <i>volucens</i> <i>dm_198</i>	4	-4	52	20	52	17	58	27	50	27	60	60	30	49	42	65	33
<i>Desulfohalobium</i> <i>actiphilum</i> <i>sp1</i>	4	-28	57	21	62	-23	58	27	60	-28	56	60	30	49	42	65	33
<i>Thermotoga</i> <i>potera</i> <i>dm_12270</i>	4	18	60	23	56	-14	53	34	69	15	65	60	30	49	42	65	33
<i>Desulfohalobium</i> <i>lufuense</i> <i>ys1</i>	4	7	62	23	62	23	52	24	59	17	56	60	30	49	42	65	33
<i>Thermotoga</i> <i>lufuensis</i> <i>dm_17291</i>	4	16	62	21	64	-3	61	24	64	-3	61	60	30	49	42	65	33
<i>Orietta tsutsugamushi</i> <i>str. borovog</i>	4	0	73	5	67	-15	64	22	67	-13	67	60	30	49	42	65	33
<i>Nostoc</i> <i>sp. pcc_7524</i>	4	-21	58	18	60	-26	68	21	61	-31	63	60	30	49	42	65	33
<i>Synechococcus</i> <i>sp. pcc_733-ab</i>	4	16	51	13	48	-4	56	20	46	30	60	60	30	49	42	65	

Таблица 1. (Конец).

Organism	Index	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
amycolatopsis_mediterranei_rb	5	19	58	65	42	62	41	65	41	66	40
rhodospirillum_rubrum_atcc_11170	5	0	53	47	42	50	47	47	39	66	40
streptomyces_violaceusniger_tu_4113	5	48	53	50	44	49	47	65	42	66	45
streptomyces_hygroscopicus_subsp_jinggangensis	5	21	58	60	33	63	31	56	43	65	35
alstipes_shahii_wal_8301	5	15	44	45	47	39	40	54	39	65	36
butyrate-producing_bacterium_sm41	5	54	46	19	54	42	49	29	49	65	51
clostridium_sp_sy8519	5	-23	49	51	42	58	35	54	43	64	36
streptomyces_avermitilis_ma-4680 = nbrc_14893	5	39	54	56	34	56	40	62	36	64	37
thiomonas_arsenitoxydans	5	3	54	53	42	56	38	57	42	64	38
thiomonas_intermedia_k12	5	-8	56	56	40	58	40	58	39	64	41
wolinella_succinogenes_dsm_1740	5	30	51	40	51	55	39	60	47	63	38
streptomyces_davawensis_jcm_4913	5	42	51	46	46	48	50	61	43	63	44
streptomyces_scabiei_87_22	5	35	56	57	42	56	43	60	43	63	44
kineococcus_radiotolerans_srs30216 = atcc_baa-1	5	-57	39	60	35	55	40	61	36	62	38
chromohalobacter_salexigens_dsm_3043	5	-8	53	56	44	51	38	38	49	61	37
xylella_fastidiosa_9a5c	5	-37	51	18	52	37	52	-32	50	61	37
chlorobium_luteolum_dsm_273	5	38	48	43	54	44	52	53	46	61	41
meiothermus_ruber_dsm_1279	5	-4	51	53	40	53	43	59	36	61	44
nitrosomonas_eutropha_c91	5	-7	65	0	76	55	51	-7	73	61	52
halothiobacillus_neapolitanus_c2	5	32	51	21	58	29	59	23	55	61	57
alstipes_finegoldii_dsm_17242	5	1	42	51	41	36	42	51	38	60	33
meiothermus_ruber_dsm_1279	5	-5	51	50	40	52	41	56	38	60	42
rhodobacter_sphaeroides_atcc_17029	5	-17	60	58	40	56	43	50	41	60	42
lactobacillus_fermentum_cect_5716	5	8	53	-15	53	44	34	-20	51	60	43
amycolatopsis_mediterranei_u32	5	16	59	60	46	58	45	59	48	60	45
desulfovibrio_gigas_dsm_1382 = atcc_19364	5	24	53	47	45	47	43	56	41	59	41
nocardiopepsis_alba_atcc_baa-2165	5	48	50	41	50	40	48	57	42	59	44
jannaschia_sp_ccs1	5	34	48	34	51	35	48	34	48	58	45
rhodopseudomonas_palustris_bisb18	5	33	51	37	49	36	50	57	43	58	46
geobacter_loveyi_sz	5	-9	59	44	55	45	44	45	52	57	45
streptomyces_bingchengensis_bcw-1	5	37	51	35	49	36	52	55	47	57	52
sulfuricella_denitrificans_skb26	5	20	54	38	49	34	49	52	45	56	46
paracoccus_denitrificans_pd1222	5	-12	48	42	41	36	45	53	41	55	43
xylella_fastidiosa_temecula1	5	-50	46	13	55	41	58	31	41	55	46
roseobacter_denitrificans_och_114	5	51	49	38	52	29	51	24	49	55	50
xylella_fastidiosa_m23	5	26	33	11	55	41	57	30	39	54	45
geobacter_sp_m21	5	17	58	45	43	42	45	50	43	53	44
fibrella_aestuarina_buz_2	5	-16	56	37	52	50	44	39	51	53	47
prevotella_dentalis_dsm_3688	5	-41	43	43	45	45	39	5	51	52	41
magnetococcus_marinus_mc-1	5	-5	42	46	49	40	42	39	46	52	46
eubacterium_limosum_kist612	5	-36	37	21	56	39	39	-9	45	51	32
starkeya_novella_dsm_506	5	31	55	32	48	38	48	42	43	51	43
geobacter_bemidjensis_bem	5	30	61	37	42	38	47	49	40	51	44
alicyclobacillus_acidocaldarius_subsp_acidocaldarii	5	35	47	15	46	22	49	43	45	51	46
dinoroseobacter_shibae_dfl_12 = dsm_16493	5	16	58	45	42	41	46	48	44	50	48
acidithiobacillus_ferrovirans_ss3	5	0	62	39	51	37	47	42	56	50	49
polymorphum_givum_sl003b-26a1	5	10	51	37	45	34	51	47	43	49	44
xylella_fastidiosa_subsp_fastidiosa_gb514	5	18	34	10	53	40	56	22	42	49	45
xylella_fastidiosa_m12	5	17	38	4	56	37	58	20	43	48	47
stackebrandtia_nassauensis_dsm_44728	5	28	59	36	53	33	55	42	52	48	54
geobacter_daltonii_frc-32	5	-27	55	30	57	44	43	22	60	47	46
spirochaeta_africana_dsm_8902	5	-41	40	46	45	43	44	40	53	47	49
chlorobium_phaeobacteroides_dsm_266	5	37	53	25	58	29	53	16	58	47	51
chlorobium_limicola_dsm_245	5	4	63	46	57	40	54	40	54	47	53
acidithiobacillus_ferroxidans_atcc_23270	5	4	58	42	46	36	48	37	53	45	48
geitlerinema_sp_pcc_7407	5	-48	40	34	50	42	43	-9	60	45	51
echinicola_vietnamensis_dsm_17526	5	22	65	6	65	27	60	5	65	45	52
azorhizobium_caulinodans_ors_571	5	43	55	30	53	32	55	43	48	45	53
xanthobacter_autotrophicus_py2	5	38	56	22	54	23	54	44	52	45	54
geobacillus_sp_jf8	5	30	55	13	56	27	45	25	56	44	48
moorella_thermoacetica_atcc_39073	5	-4	50	37	44	32	44	42	52	44	52
prosthecochloris_aestuarii_dsm_271	5	32	51	19	61	35	57	17	57	44	57
candidatus_tremblaya_phenacola_pave	5	38	51	9	56	39	45	25	50	43	45
symbiobacterium_thermophilum_iam_14863	5	-12	61	38	46	38	45	35	51	40	46
chloroherpeton_thalassium_atcc_35110	5	-59	53	39	60	31	53	24	62	39	51
desulfurispirillum_indicum_s5	5	-2	51	27	59	32	49	25	62	36	55
thermobifida_fusca_yx	5	10	54	26	52	28	54	22	51	35	53
campylobacter_curvus_525_92	5	2	59	31	56	29	56	22	60	33	53
candidatus_uzinura_diaspidicola_str_asner	5	23	50	20	52	27	56	30	53	32	51
chlorobium_phaeovibrioides_dsm_265	5	8	50	22	60	27	63	24	55	32	60
desulfohalobium_retbaense_dsm_5692	5	3	56	23	61	27	55	21	60	31	59
mycobacterium_kansasii_atcc_12478	5	-25	51	21	51	14	57	19	55	31	63
runella_slithyiformis_dsm_19594	5	-17	59	4	67	24	55	-2	62	30	54
sulfobacillus_acidophilus_tpy	5	21	58	10	58	16	52	-12	56	29	55
mycobacterium_marinum_m	5	-38	54	16	58	12	57	23	65	27	63
kyrpidia_tusciae_dsm_2912	5	3	50	16	59	15	58	19	57	26	59
chlorobium_phaeobacteroides_bs1	5	6	61	24	64	22	59	24	65	25	60
caldiilinea_aerophila_dsm_14535 = nbrc_104270	5	18	50	10	59	13	56	3	53	24	59
nitrosococcus_watsonii_c-113	5	-32	52	17	57	17	56	-16	55	24	62
roseiflexus_castenholzii_dsm_13941	5	21	59	11	57	7	54	-3	61	22	53
candidatus_hodgkinea_cicadicola_dsem	5	0	49	18	37	19	46	14	47	20	50
mycobacterium_tuberculosis_casnitr204	5	-25	54	14	54	4	58	10	59	18	65

Таблица 1. Значения типов индекса EEI для 2582 организмов из домена Бактерии (Конец).

M1-M5 – значения средних положений (рангов) генов рибосомных белков в отсортированных по увеличению индекса EEI списках генов организмов, R1-R5 – стандартные отклонения от средних (см. главу «Методы и алгоритмы»). Синим цветом выделены наибольшие значения параметра M для организма, красным – организмы с максимальным параметром M < 30.

Таблица 2. Значения типов индекса ЕЕІ для 165 организмов из домена Археи.

M1-M5 – значения средних положений (рангов) рибосомных генов в отсортированных по увеличению индекса ЕЕІ списках генов организмов, R1-R5 – стандартные отклонения от средних (см. главу «Методы и алгоритмы»). Синим цветом выделены наибольшие значения параметра М для организма, красным – организмы с максимальным М<30.

Organism	Index	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
methanococcus_maripaludis_c6	1	92	5	63	45	-18	62	92	10	-19	59
thermococcus_kodakarensis_kod1	1	88	21	-10	67	-55	43	63	45	9	69
thermococcus_omnirivus_na1	1	88	21	-22	65	-61	43	64	48	-4	74
thermococcus_sp_4557	1	84	34	-12	65	-47	51	30	64	-26	63
thermococcus_sp_c11	1	77	40	6	63	-37	56	58	54	-1	69
thermococcus_sp_am4	1	76	37	7	62	-37	55	55	54	0	68
thermococcus_gammatorans_ej3	1	75	36	0	65	-35	55	61	48	4	67
halogeometricum_borinquense_dsm_11551	1	73	44	-3	51	-26	57	60	49	11	66
methanocella_arvoryzae_mre50	1	72	34	39	54	2	59	72	37	41	59
haloarcuula_marismortui_atcc_43049	1	67	49	-20	54	-28	55	40	52	2	64
pyrococcus_yayanosii_ch1	1	66	46	8	63	-26	64	54	51	7	71
haloarcuula_hispanica_atcc_33960	1	62	48	-10	55	-24	53	53	46	12	63
haloarcuula_hispanica_n601	1	62	49	-7	54	-22	52	55	42	14	62
thermococcus_litoralis_dsm_5473	1	60	61	38	58	-12	62	57	56	14	65
natronomonas_pharaonis_dsm_2160	1	59	54	0	59	-9	59	49	49	23	61
halophilic_archaeon_d131	1	58	54	-18	64	-22	67	22	55	2	68
archaeoglobus_veneficus_snp6	1	57	47	31	61	-27	64	54	46	2	71
natrialba_magadii_atcc_43099	1	56	50	2	50	-19	57	46	52	0	67
methanoregula_boonei_6a8	1	51	50	23	57	6	56	33	56	20	60
methanoregula_formicica_smsp	1	49	42	25	56	-11	50	39	53	13	61
halorhabdus_tiamatea_sar4b	1	49	51	-1	56	-23	59	33	52	-4	64
halorhabdus_utahensis_dsm_12940	1	47	48	-5	54	-23	58	40	54	5	66
salinarchaeum_sp_harcht-bsk1	1	46	51	6	56	-13	60	41	50	6	64
candidatus_caldiarchaeum_subterraneum	1	39	50	28	59	-1	57	37	50	15	65
vulcanisaeta_distributa_dsm_14429	1	38	67	24	58	-18	62	36	60	-3	68
metallophera_cuprina_ar-4	1	37	62	22	62	2	63	25	64	9	64
haloquadratum_walsbyi_dsm_16790	1	37	63	-1	61	-31	59	14	63	-15	63
haloquadratum_walsbyi_c23	1	31	67	-10	65	-42	59	9	69	-25	66
halalkalicoccus_jeotgali_b3	1	29	55	11	54	-6	52	5	48	10	52
picropilus_torridus_dsm_9790	2	-41	57	78	26	-36	63	34	63	-50	53
methanothermus_fervidus_dsm_2088	2	-3	67	74	30	1	62	46	59	-8	61
methanobacterium_sp_al-21	2	-36	54	70	44	-14	63	65	44	-30	55
caldisphaera_lagunensis_dsm_15908	2	-22	58	69	35	-24	61	37	63	-32	57
methanocaldococcus_vulcanius_m7	2	11	53	69	40	-14	57	68	39	-7	55
methanosphaera_stadtmanae_dsm_3091	2	-15	57	67	29	-20	57	67	38	-34	48
nanoarchaeum_equitans_kin4-m	2	3	58	65	26	-21	65	53	42	-23	62
thermoplasma_volcanium_gss1	2	5	68	62	46	-7	64	44	59	-11	64
ferroplasma_acidarmanus_fer1	2	2	53	61	44	-17	58	51	54	-16	52
fervidococcus_fontis_kam940	2	17	52	58	34	12	59	53	55	12	54
staphylothermus_marinus_f1	2	-6	65	58	42	-8	70	38	60	-18	63
methanosphaerula_palustris_e1-9c	2	-23	41	57	39	32	46	46	44	29	48
thermogladus_cellulolyticus_1633	2	-29	39	57	46	23	54	22	54	7	56
thermoproteus_uzoniensis_768-20	2	-11	53	56	41	29	50	33	59	23	57
thermoproteus_tenax_kra_1	2	-3	51	56	42	32	50	47	53	35	57
staphylothermus_hellenicus_dsm_12710	2	9	63	55	43	-15	62	42	56	-23	63
thermofilum_pendens_hrk_5	2	-63	28	55	51	51	49	17	61	28	52
aciduliprofundum_boonei_1469	2	-24	61	53	45	-27	62	33	53	-38	57
desulfurococcus_fermentans_dsm_16532	2	31	63	52	50	-2	65	52	55	3	68
thermoplasma_acidophilum_dsm_1728	2	27	59	52	52	-4	56	43	54	7	68
sulfolobus_islandicus_m_16_27	2	17	59	52	54	3	66	50	58	6	59
methanohalobium_evestigatum_z-7303	2	-17	57	51	46	-16	56	47	47	-21	54
desulfurococcus_kamchatkensis_1221n	2	29	61	51	49	0	59	49	54	13	62
aciduliprofundum_sp_mar08-339	2	5	61	48	40	-21	51	32	55	-28	52
sulfolobus_islandicus_l_s_2_15	2	16	61	48	55	5	66	44	62	7	59
methanoculleus_marisnigri_jr1	2	-19	48	47	38	20	49	45	44	23	55
acidilobus_saccharovorans_345-15	2	-42	30	46	44	36	46	-19	48	-4	47
methanosalsum_zhilinae_dsm_4017	2	0	62	46	56	4	56	36	56	-7	56
desulfurococcus_mucosus_dsm_2162	2	22	56	42	56	9	63	41	56	19	64
methanosarcina_barkeri_str_fusaro	2	-53	45	40	51	-32	60	23	57	-51	45
pyrobaculum_arsenicum_dsm_13514	2	-22	57	40	53	30	53	19	62	20	57
pyrobaculum_caldifontis_jcm_11548	2	-26	48	39	43	35	49	22	61	26	57
haloferax_volcanii_ds2	2	-15	52	38	45	29	48	35	45	33	45
methanoculleus_bourgenensis_ms2	2	-25	39	38	50	23	50	23	52	23	55
pyrobaculum_sp_1860	2	-1	46	34	45	19	47	29	53	33	58
halobacterium_sp_nrc-1	2	-22	45	33	40	22	44	31	44	27	43
halobacterium_salinarum_r1	2	-24	44	33	40	23	43	27	45	25	42
pyrobaculum_oguniense_te7	2	-19	53	32	51	25	49	26	54	26	53
pyrobaculum_aerophilum_str_im2	2	4	64	32	59	14	53	25	62	16	63
methanococcus_maripaludis_c7	4	93	6	64	43	-23	63	94	6	-29	55
methanococcus_maripaludis_s2	4	90	16	64	45	-30	59	92	14	-29	53
methanococcus_maripaludis_c5	4	89	24	64	44	-21	62	90	24	-27	55
methanobrevibacter_ruminantium_m1	4	10	51	42	48	-15	61	89	7	-10	56
methanococcus_voltae_a3	4	-32	45	69	33	-37	51	89	19	-43	45
methanocaldococcus_fervens_ag86	4	9	58	75	34	-27	65	89	23	-22	63
methanococcus_vannielii_sb	4	-42	55	72	30	-35	62	87	21	-49	49
methanobrevibacter_sp_abm4	4	58	40	26	51	-19	64	86	13	-1	64
methanococcus_maripaludis_x1	4	82	32	58	51	-26	60	83	34	-26	55
methanocaldococcus_jannaschii_dsm_2661	4	-5	60	76	34	-18	64	83	34	-18	63
methanobrevibacter_smithii_atcc_35061	4	38	45	41	49	-32	59	82	20	-19	62
methanotorris_igneus_kol_5	4	-4	49	66	39	-4	62	82	28	-9	57
methanobacterium_sp_swam-1	4	3	48	62	38	-5	60	78	28	-6	49
methanocaldococcus_sp_fs406-22	4	13	56	75	37	-11	69	78	35	-10	66
thermoplasmatales_archaeon_brna1	4	62	38	46	42	22	47	77	25	57	42

Таблица 2. (Конец).

Organism	Index	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
<i>methanococcus aeolicus_nankai-3</i>	4	-27	51	68	37	-29	59	77	31	-33	55
<i>methanocorpusculum labreanum_z</i>	4	73	47	37	52	5	52	75	44	59	50
<i>methanobacterium_sp_mb1</i>	4	-32	62	73	34	-16	61	74	33	-29	52
<i>candidatus_methanomethylophilus_alvus_mx1201</i>	4	64	36	49	52	10	52	74	37	43	55
<i>methanocaldococcus infernus_me</i>	4	-4	58	70	40	0	58	73	34	-5	55
<i>candidatus_nitrosopumilus_sp_ar2</i>	4	3	59	65	43	-1	58	73	40	-3	57
<i>methanothermococcus okinawensis_ih1</i>	4	-13	53	60	46	-34	61	72	42	-38	52
<i>methanobacterium_psychrophilum_r15</i>	4	67	37	52	44	-17	57	71	36	-26	48
<i>methanospirillum hungatei_jf1</i>	4	59	49	39	52	-37	56	70	38	-63	33
<i>ferroglobus placidus_dsm_10642</i>	4	42	50	60	42	-16	59	69	36	-5	60
<i>methanococcoides burtonii_dsm_6242</i>	4	68	44	32	57	-23	58	69	43	-30	50
<i>thermococcus barophilus_mp</i>	4	62	56	44	53	-24	60	69	44	-5	68
<i>methanosarcina mazei_go1</i>	4	63	45	44	41	-30	53	68	36	-48	48
<i>nitrosopumilus maritimus_scml</i>	4	13	65	66	42	-1	61	67	49	-4	58
<i>archaeoglobus sulfatocalidus_pm70-1</i>	4	41	58	58	41	-5	57	65	35	13	59
<i>methanothermobacter marburgensis_str_marburg</i>	4	52	54	47	52	2	57	64	39	23	55
<i>haloferax mediterranei_atcc_33500</i>	4	58	43	14	52	-15	54	64	45	24	56
<i>natrinema_sp_j7-2</i>	4	47	47	17	43	-5	49	63	33	28	53
<i>methanosacta harundinacea_6ac</i>	4	-6	42	57	31	40	36	63	37	56	39
<i>methanopyrus kandleri_av19</i>	4	21	50	39	45	19	56	63	39	39	53
<i>methanosarcina mazei_tuc01</i>	4	62	45	35	45	-24	54	63	41	-27	62
<i>pyrococcus furiosus_com1</i>	4	32	59	58	44	-14	59	63	45	-11	61
<i>methanosarcina acetivorans_c2a</i>	4	59	52	38	52	-32	55	63	49	-46	50
<i>methanothermobacter thermotrophicus_str_delta</i>	4	47	57	51	47	25	55	62	43	40	55
<i>candidatus_nitrosopumilus koreensis_ar1</i>	4	10	57	57	41	-14	52	62	45	-12	53
<i>sulfolobus tokodaii_str_7</i>	4	40	51	45	43	19	60	62	46	30	58
<i>pyrococcus furiosus_dsm_3638</i>	4	37	60	61	41	-14	59	62	48	-3	65
<i>archaeoglobus profundus_dsm_5631</i>	4	40	48	52	43	-8	62	61	38	10	63
<i>methanocella conradii_hz254</i>	4	56	43	23	59	-22	57	61	46	18	67
<i>methanoplanus petrolearius_dsm_11571</i>	4	35	46	47	55	-10	58	61	50	-22	48
<i>archaeoglobus fulgidus_dsm_4304</i>	4	51	45	36	49	3	60	60	36	25	61
<i>sulfolobus islandicus_lal141</i>	4	27	56	56	49	3	65	59	46	6	59
<i>methanocella paludicola_sanae</i>	4	54	47	31	60	-8	54	59	48	19	64
<i>candidatus_methanomassiliococcus intestinalis_issouir</i>	4	-39	51	51	49	-31	52	59	53	-44	44
<i>halopiger xanaduensis_sh-6</i>	4	25	56	38	44	5	55	58	36	15	56
<i>sulfolobus solfataricus_p2</i>	4	36	53	43	51	13	63	58	45	22	62
<i>hyperthermus butylicus_dsm_5456</i>	4	42	58	35	60	22	58	58	48	49	56
<i>pyrococcus horikoshii_ot3</i>	4	29	63	56	46	-5	66	57	42	-5	63
<i>acidianus hospitalis_w1</i>	4	21	62	53	44	10	65	57	48	12	64
<i>ignicoccus hospitalis_kin4i</i>	4	32	47	36	58	8	60	56	44	29	57
<i>sulfolobus islandicus_hve104</i>	4	24	56	54	51	9	66	56	48	7	59
<i>pyrococcus abyssi_ge5</i>	4	46	58	47	54	-18	68	56	49	-9	71
<i>natronobacterium gregoryi_sp2</i>	4	39	50	19	52	-9	56	56	49	15	61
<i>sulfolobus islandicus_rey15a</i>	4	26	55	51	53	9	65	56	49	7	59
<i>halorubrum lacusprofundi_atcc_49239</i>	4	30	47	15	44	-2	48	55	38	31	56
<i>natronococcus occultus_sp4</i>	4	12	57	28	48	2	55	55	42	15	56
<i>natrinema pellirubrum_dsm_15624</i>	4	26	60	34	48	11	52	55	46	26	54
<i>sulfolobus islandicus_m_14_25</i>	4	19	59	50	55	4	66	55	52	7	60
<i>thermofilum_sp_1910b</i>	4	52	49	35	48	-28	59	54	44	-4	66
<i>sulfolobus islandicus_y_g_57_14</i>	4	25	57	47	52	8	63	54	52	11	55
<i>aeropyrum camini_sy1_jcm_12091</i>	4	30	48	30	55	16	56	53	45	44	51
<i>sulfolobus acidocaldarius_dsm_639</i>	4	34	58	40	51	3	63	53	46	9	63
<i>pyrolobus fumarii_la</i>	4	47	48	15	56	-3	54	53	57	43	54
<i>cenarchaeum symbiosum_a</i>	4	37	38	32	54	4	52	52	38	36	45
<i>natronomonas moolapensis_8_8_11</i>	4	27	52	14	48	3	49	52	40	36	50
<i>candidatus_nitrososphaera gargensis_ga9_2</i>	4	47	44	29	57	-29	51	52	45	8	63
<i>methanohalophilus mahii_dsm_5219</i>	4	43	54	29	54	-14	57	52	46	-20	51
<i>sulfolobus solfataricus_982</i>	4	25	57	46	50	1	66	52	51	9	64
<i>sulfolobus islandicus_m_16_4</i>	4	21	55	50	54	5	64	52	52	4	59
<i>sulfolobus acidocaldarius_susaz</i>	4	28	65	46	48	2	64	52	54	9	60
<i>halomicrobium mukohataei_dsm_12286</i>	4	37	55	28	48	10	53	51	42	25	57
<i>aeropyrum permix_k1</i>	4	32	46	23	60	4	58	51	45	32	62
<i>sulfolobus acidocaldarius_rm12i</i>	4	35	62	35	51	2	62	51	52	14	62
<i>thermococcus sibiricus_mm_739</i>	4	42	60	35	58	-14	63	51	57	5	64
<i>methanosacta concilia_gp6</i>	4	31	48	37	52	-16	52	50	46	9	63
<i>methanosacta thermophila_pt</i>	4	26	56	45	51	-11	63	50	46	-2	69
<i>haloterrigena turkmenica_dsm_5511</i>	4	11	65	39	53	15	59	50	46	23	59
<i>caldivirga maquilgensis_ic-167</i>	4	35	63	40	49	-12	63	50	47	-6	69
<i>sulfolobus acidocaldarius_n8</i>	4	33	59	35	51	2	62	50	51	13	63
<i>sulfolobus islandicus_y_n_15_51</i>	4	22	55	47	54	3	66	50	53	4	60
<i>sulfolobus islandicus_l_d_8_5</i>	4	22	58	45	56	4	64	50	54	5	58
<i>thermosphaera aggregans_dsm_11486</i>	4	46	60	31	64	-30	63	49	56	-22	65
<i>pyrococcus_sp_na2</i>	4	38	65	47	54	-7	67	49	56	1	69
<i>ignisphaera aggregans_dsm_17230</i>	4	15	55	46	40	3	67	48	45	2	63
<i>methanomethylovorans hollandica_dsm_15978</i>	4	36	55	43	58	-12	52	47	53	-23	51
<i>pyrococcus_sp_st04</i>	4	36	63	43	58	-21	65	47	59	-18	70
<i>candidatus_korarchaeum cryptofilum_op8</i>	4	28	56	41	58	3	66	46	54	17	64
<i>vulcanisaeta moutnovskia_768-28</i>	4	33	66	39	47	-1	59	43	54	8	65
<i>halovivax ruber_xh-70</i>	4	30	50	17	52	0	53	42	46	23	53
<i>pyrobaculum islandicum_dsm_4184</i>	4	-24	41	39	48	35	45	42	58	37	51
<i>pyrobaculum neutrophilum_v24sta</i>	4	-10	46	35	51	25	56	38	51	28	57
<i>metalkosphaera sedula_dsm_5348</i>	4	16	69	13	67	5	64	22	67	7	71



**Таблица 3. Значения типов индекса EEI для 73 особенных организмов со значением параметра M < 30.**

M1-M5 – значения средних положений (рангов) генов рибосомных белков в отсортированных по увеличению индекса EEI списках генов организмов, R1-R5 – стандартные отклонения от средних (см. главу «Методы и алгоритмы»). Синим цветом выделены наибольшие значения параметра M для организма.

Organism	Domain	Index	M1	R1	M2	R2	M3	R3	M4	R4	M5	R5
halalkalicoccus_jeotgali_b3	Archaea	1	29	55	11	54	-6	52	5	48	10	52
bacillus_coagulans_36d1	Bacteria	1	28	55	17	54	-1	48	25	54	13	58
synechococcus_sp_pcc_6312	Bacteria	1	28	57	14	59	-18	61	8	55	-25	59
bacillus_coagulans_2-6	Bacteria	1	27	55	20	52	5	48	22	52	11	55
helicobacter_felis_atcc_49179	Bacteria	1	25	64	14	60	8	56	20	62	14	57
encephalitozoon_cuniculi_gb-m1	Eukaryota	1	22	51	14	62	-14	61	19	51	-9	58
roseiflexus_sp_rs-1	Bacteria	1	23	59	4	58	-2	59	9	56	13	64
wolbachia_endosymbiont_of_culex_quinquefasciatus	Bacteria	1	21	55	10	59	11	58	20	49	19	55
jonesia_dentrificans_dsm_20603	Bacteria	1	15	32	-38	53	-15	62	-16	52	0	57
bacillus_selenitireducens_mls10	Bacteria	1	15	54	15	64	4	56	6	67	3	52
mycoplasma_haemocanis_str_illinois	Bacteria	1	-4	66	-6	63	-6	70	-19	55	-16	65
candidatus_sucia_muelleri_gwss	Bacteria	2	-12	45	29	40	25	58	23	58	4	51
nitrosococcus_oceani_atcc_19707	Bacteria	2	-30	54	29	56	18	56	-10	52	22	58
helicobacter_bizzozeronii_ciii-1	Bacteria	2	12	63	29	62	1	58	24	58	2	62
haliscomenobacter_hydrossis_dsm_1100	Bacteria	2	-6	57	28	58	0	57	26	59	-4	52
candidatus_blochmannia_pennsylvanicus_str_bpen	Bacteria	2	-16	64	28	58	-8	69	13	59	-20	67
mycoplasma_suis_ki3806	Bacteria	2	18	49	27	45	0	50	23	37	3	46
candidatus_blochmannia_chromaiodes_str_640	Bacteria	2	-18	62	27	57	-2	69	9	63	-18	66
mycoplasma_suis_str_illinois	Bacteria	2	6	48	26	46	0	49	17	39	2	46
halothece_sp_pcc_7418	Bacteria	2	-5	59	26	60	-36	62	22	58	-34	59
desulfobacca_acetoxidans_dsm_11109	Bacteria	2	-2	54	26	62	19	58	15	59	25	59
candidatus_sucia_muelleri_dmin	Bacteria	2	-9	46	25	40	21	58	4	61	4	51
candidatus_carsonella_ruddii_pv	Bacteria	2	-17	42	25	41	14	56	4	55	-8	44
thermobacillus_composti_kwc4	Bacteria	2	2	50	25	46	21	46	19	49	23	45
candidatus_mycoplasma_haemominutum_birmingham	Bacteria	2	-7	53	24	48	2	52	11	57	-8	52
candidatus_carsonella_ruddii_dc	Bacteria	2	-16	47	23	43	1	55	5	58	-12	45
rivularia_sp_pcc_7116	Bacteria	2	-27	56	23	58	-47	57	1	57	-53	52
pseudanabaena_sp_pcc_7367	Bacteria	2	-9	57	21	60	-3	67	-7	57	-5	64
dactylococcopsis_salina_pcc_8305	Bacteria	2	-33	54	21	61	-43	57	-2	60	-47	51
thermomicrobium_roseum_dsm_5159	Bacteria	2	-3	50	20	51	4	56	9	48	8	58
candidatus_nasuia_deltoccephalinicola_str_nas-alf	Bacteria	2	-8	39	19	28	-2	48	-4	43	-9	45
sphaerobacter_thermophilus_dsm_20745	Bacteria	2	9	57	18	51	13	57	16	52	18	59
gloeocapsa_sp_pcc_7428	Bacteria	2	2	57	15	59	-44	51	-10	56	-49	46
oscillatoria_acuminata_pcc_6304	Bacteria	2	9	56	14	56	-19	55	12	57	-28	52
prochlorococcus_marinus_str_mit_9313	Bacteria	2	-12	50	5	61	-26	58	-16	50	-24	56
candidatus_chloracidobacterium_thermophilum_b	Bacteria	3	-25	55	5	60	28	56	-15	53	11	54
candidatus_sucia_muelleri_str_sulcia-alf	Bacteria	3	6	48	18	44	27	47	21	48	18	46
candidatus_carsonella_ruddii_hc_isolate_thao2000	Bacteria	3	-2	49	13	37	26	53	9	57	17	52
candidatus_sucia_muelleri_smdsem	Bacteria	3	0	46	24	50	25	59	20	52	12	54
candidatus_carsonella_ruddii_pc_isolate_nhv	Bacteria	3	6	45	13	39	23	47	20	52	9	47
gamma_proteobacterium_hdn1	Bacteria	3	-8	71	7	61	22	62	-6	57	16	58
desulfitobacterium_hafniense_dcb-2	Bacteria	3	16	67	16	62	18	55	12	60	12	51
candidatus_carsonella_ruddii_ht_isolate_thao2000	Bacteria	3	1	47	16	41	17	50	16	56	6	48
candidatus_carsonella_ruddii_ce_isolate_thao2000	Bacteria	3	0	52	13	39	16	50	7	59	-2	45
sulfobacillus_acidophilus_dsm_10332	Bacteria	3	8	65	10	60	16	55	-23	57	15	64
mycoplasma_haemofelis_ohio2	Bacteria	3	-16	56	-21	59	7	67	-36	45	-7	66
mycoplasma_haemofelis_str_langford_1	Bacteria	3	-17	57	-16	60	3	69	-28	52	-1	69
candidatus_tremblaya_princeps_pcit	Bacteria	3	-24	37	-9	35	0	39	-3	35	-9	32
candidatus_tremblaya_princeps_pcval	Bacteria	3	-30	27	-10	24	-1	30	-17	16	-15	19
ehrlichia_muris_as145	Bacteria	4	14	59	26	61	2	60	29	54	10	59
thermincola_potens_jr	Bacteria	4	19	61	24	60	18	61	29	58	18	60
candidatus_methylomirabilis_oxifera	Bacteria	4	5	39	21	56	2	58	28	53	19	59
helicobacter_heilmannii_asb1_4	Bacteria	4	20	60	16	52	-17	49	28	58	7	60
ehrlichia_camis_str_jake	Bacteria	4	14	63	21	60	12	61	28	59	12	66
thiocystis_violascens_dsm_198	Bacteria	4	-4	52	20	52	17	58	27	50	27	60
desulfosporosinus_acidiphilus_sj4	Bacteria	4	-28	57	21	62	-23	58	27	60	-28	56
thermactinogenium_phaeum_dsm_12270	Bacteria	4	10	49	23	56	-4	52	26	58	-3	55
desulfitobacterium_hafniense_y51	Bacteria	4	7	62	23	62	23	52	24	59	17	56
thermovirga_lienii_dsm_17291	Bacteria	4	16	62	21	64	-3	61	24	64	-3	61
metallophaera_sedula_dsm_5348	Archaea	4	16	69	13	67	5	64	22	67	7	71
orientia_tsutsugamushi_str_boryong	Bacteria	4	0	73	5	67	-15	64	22	67	-13	67
nostoc_sp_pcc_7524	Bacteria	4	-21	58	18	60	-26	68	21	61	-31	63
synechococcus_sp_ja-3-3ab	Bacteria	4	16	51	13	48	-4	58	20	46	1	61
candidatus_carsonella_ruddii_cs_isolate_thao2000	Bacteria	4	0	48	10	39	7	59	13	57	-1	48
sulfobacillus_acidophilus_tpy	Bacteria	5	21	58	10	58	16	52	-12	56	29	55
mycobacterium_marinum_m	Bacteria	5	-38	54	16	58	12	57	23	65	27	63
kyrpidia_tusciae_dsm_2912	Bacteria	5	3	50	16	59	15	58	19	57	26	59
chlorobium_phaeobacteroides_bs1	Bacteria	5	6	61	24	64	22	59	24	65	25	60
cauldilinea_aerophila_dsm_14535_nbrc_104270	Bacteria	5	18	50	10	59	13	56	3	53	24	59
nitrosococcus_watsonii_c-113	Bacteria	5	-32	52	17	57	17	56	-16	55	24	62
roseiflexus_castenholzii_dsm_13941	Bacteria	5	21	59	11	57	7	54	-3	61	22	53
candidatus_hodgkinia_cicadicola_dsem	Bacteria	5	0	49	18	37	19	46	14	47	20	50
mycobacterium_tuberculosis_casnitr204	Bacteria	5	-25	54	14	54	4	58	10	59	18	65

Таблица 4. Типы индексов для 62 штаммов *Mycoplasma*.

Синим цветом в строке выделен тот тип индекса EEI, который работает в соответствующем штамме (наибольшее значение параметра M в строке). Красным выделены штаммы с наименьшим смещением генов рибосомных белков в сторону высокоэкспрессирующихся генов ( $M \leq 30$ ). Для штаммов с одинаковыми значениями параметра M для разных типов индекса в скобках дополнительно приведены значения параметра R.

Strain	Index	M1 (R1)	M2 (R2)	M3 (R3)	M4 (R4)	M5 (R5)
<i>Mycoplasma wenyonii</i> str. Massachusetts	1	70	-9	-21	69	25
<i>Candidatus Mycoplasma haemolamae</i> str. Purdue	1	30	5	20	25	29
<i>Mycoplasma fermentans</i> M64	2	-57	79	-50	67	-74
<i>Mycoplasma hyorhinitis</i> MCLD	2	-20	79	-23	45	-42
<i>Mycoplasma fermentans</i> PG18	2	-53	78	-50	75	-66
<i>Mycoplasma hyorhinitis</i> HUB 1	2	-9	78	-16	46	-29
<i>Mycoplasma capricolum</i> subsp. capricolum ATCC 27343	2	-60	78	-26	40	-52
<i>Mycoplasma fermentans</i> JER	2	-40	77	-42	71	-62
<i>Mycoplasma hyorhinitis</i> SK76	2	-15	77	-24	48	-40
<i>Mycoplasma hyopneumoniae</i> 232	2	-61	77	-24	-17	-57
<i>Mycoplasma crocodyli</i> MP145	2	-44	76	-44	56	-57
<i>Mycoplasma mycoides</i> subsp. mycoides SC str. Gladysdale	2	-37	76	-24	61	-42
<i>Mycoplasma mycoides</i> subsp. mycoides SC str. PG1	2	-45	76	-21	65	-43
<i>Mycoplasma leachii</i> 990146	2	-59	76	-29	70	-49
<i>Mycoplasma mycoides</i> subsp. capri LC str. 95010	2	-53	76	-27	25	-51
<i>Mycoplasma leachii</i> PG50	2	-56	75	-31	41	-56
<i>Mycoplasma cynos</i> C142	2	-39	74	-29	47	-47
<i>Mycoplasma synoviae</i> 53	2	-37	71	-8	50	-32
<i>Mycoplasma penetrans</i> HF 2	2	-1	71	-32	60	-37
<i>Mycoplasma hyopneumoniae</i> 7448	2	-69	71	-29	-21	-63
<i>Mycoplasma putrefaciens</i> Mput9231	2	-17	70 (36)	-24	70 (49)	-37

<i>Mycoplasma hyorhinitis</i> GDL 1	2	-20	70	-23	40	-36
<i>Mycoplasma hyopneumoniae</i> J	2	-70	70	-30	-23	-62
<i>Mycoplasma mobile</i> 163K	2	-48	70	-9	53	-30
<i>Mycoplasma hominis</i> ATCC 23114	2	-7	69	-28	67	-38
<i>Mycoplasma hyopneumoniae</i> 168 L	2	-64	69	-26	-27	-58
<i>Mycoplasma putrefaciens</i> KS1	2	-20	68	-21	59	-41
<i>Mycoplasma pulmonis</i> UAB CTIP	2	-13	68	5	48	-16
<i>Mycoplasma hyopneumoniae</i> 168	2	-67	68	-28	-33	-60
<i>Mycoplasma conjunctivae</i> HRC/581	2	-8	67	-18	50	-29
<i>Mycoplasma arthritidis</i> 158L3-1	2	9	62	-23	52	-21
<i>Mycoplasma genitalium</i> G37	2	-66	59	-39	-40	-64
<i>Mycoplasma genitalium</i> M6282	2	-52	53	-24	-29	-48
<i>Mycoplasma gallisepticum</i> str. R (high)	2	18	53	2	46	-3
<i>Mycoplasma gallisepticum</i> str. R (low)	2	19	53	2	46	-2
<i>Mycoplasma genitalium</i> M2288	2	-53	52	-30	-30	-51
<i>Mycoplasma genitalium</i> M6320	2	-50	50	-26	-29	-47
<i>Mycoplasma genitalium</i> M2321	2	-51	50	-26	-30	-46
<i>Mycoplasma gallisepticum</i> NY01 2001.047-5-1P	2	24	49	-7	49	-2
<i>Mycoplasma gallisepticum</i> ca06 2006 052-5-2p	2	23	49	-7	47	-1
<i>Mycoplasma gallisepticum</i> NC96 1596-4-2P	2	24	48	-6	47	-2
<i>Mycoplasma suis</i> KI3806	2	18	27	0	23	3
<i>Mycoplasma suis</i> str. Illinois	2	6	26	0	17	2
<i>Mycoplasma pneumoniae</i> M129-B7	2	-10	25	-27	23	-30
<i>Mycoplasma pneumoniae</i> FH	2	-13	24	-32	21	-31
<i>Candidatus Mycoplasma haemominutum</i> Birmingham 1	2	-8	24	2	8	-9
<i>Mycoplasma haemocanis</i> str. Illinois	2	-15	-6 (63)	-6 (70)	-19	-24

<i>Mycoplasma haemofelis</i> Ohio2	3	-23	-21	7	-36	-3
<i>Mycoplasma haemofelis</i> str. Langford 1	3	-25	-16	3	-35	-12
<i>Mycoplasma bovis</i> Hubei-1	4	25	66	-44	80	-27
<i>Mycoplasma agalactiae</i> PG2	4	22	64	-27	73	-27
<i>Mycoplasma bovis</i> HB0801	4	15	67	-38	69	-34
<i>Mycoplasma agalactiae</i>	4	26	64	-24	66	-26
<i>Mycoplasma bovis</i> PG45	4	13	63	-40	66	-35
<i>Mycoplasma gallisepticum</i> str. F	4	31	54	1	62	3
<i>Mycoplasma gallisepticum</i> NC08 2008.031-4-3P	4	26	49	-7	53	-1
<i>Mycoplasma gallisepticum</i> WI01 2001.043-13-2P	4	26	50	-7	52	-1
<i>Mycoplasma gallisepticum</i> NC06 2006.080-5-2P	4	26	49	-7	51	-2
<i>Mycoplasma gallisepticum</i> NC95 13295-2-2P	4	26	49	-7	51	-2
<i>Mycoplasma gallisepticum</i> VA94 7994-1-7P	4	26	48	-7	51	-2
<i>Mycoplasma pneumoniae</i> 309	4	-9	23	-26	25	-29
<i>Mycoplasma pneumoniae</i> M129	4	-10	22	-25	25	-29

Таблица 4. Типы индексов для 62 штаммов *Mycoplasma* (Конец).

Синим цветом в строке выделен тот тип индекса EEI, который работает в соответствующем штамме (наибольшее значение параметра M в строке). Красным выделены штаммы с наименьшим смещением генов рибосомных белков в сторону высокоэкспрессирующихся генов ( $M \leq 30$ ). Для штаммов с одинаковыми значениями параметра M для разных типов индекса в скобках дополнительно приведены значения параметра R.

**Таблица 5. Профили средних по всем генам значений LCI индивидуальных нуклеотидов 62 исследованных микоплазм.**

Профили построены для районов старт- и стоп-кодонов генов  $\pm 300$  нуклеотидов. Середина ячейки «LCI( $i, j$ ) 5'» - начало старт-кодона, середина ячейки «LCI( $i, j$ ) 3'» - начало стоп-кодона.

Organism	Number of genes	LCI( $i, j$ ) 5'	LCI( $i, j$ ) 3'	Index	M
<i>Mycoplasma wenyonii</i> str massachusetts	652			1	70
Candidatus <i>Mycoplasma haemolamae</i> str purdue	925			1	29
<i>Mycoplasma fermentans</i> m64	1050			2	80
<i>Mycoplasma capricolum</i> subsp capricolum atcc 27343	812			2	79
<i>Mycoplasma hyorhinis</i> mcd1	778			2	79
<i>Mycoplasma hyorhinis</i> hub-1	658			2	78
<i>Mycoplasma fermentans</i> jer	797			2	78
<i>Mycoplasma fermentans</i> pg18	893			2	78
<i>Mycoplasma hyorhinis</i> gdl-1	647			2	77
<i>Mycoplasma hyopneumoniae</i> 232	691			2	77
<i>Mycoplasma hyorhinis</i> sk76	753			2	77
<i>Mycoplasma leachii</i> 990146	905			2	76
<i>Mycoplasma mycoides</i> subsp capri lc str 95010	922			2	76
<i>Mycoplasma mycoides</i> subsp mycoides sc str gladysdale	1095			2	76
<i>Mycoplasma mycoides</i> subsp mycoides sc str pgl	1017			2	76
<i>Mycoplasma crocodyli</i> mp145	689			2	76

Таблица 5. (Продолжение).


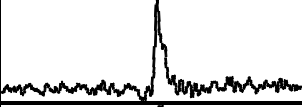

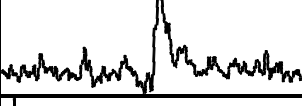








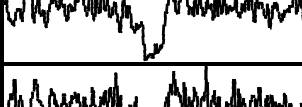
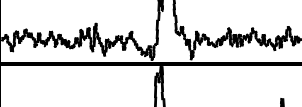
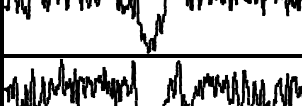

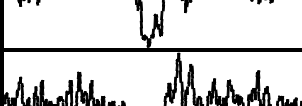
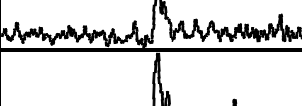



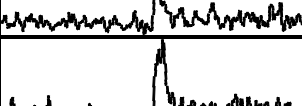

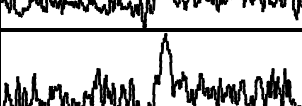




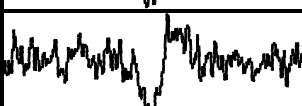
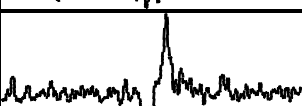


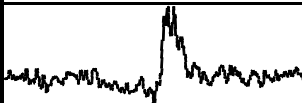
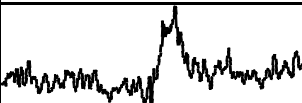



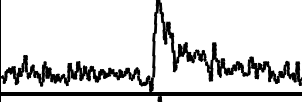










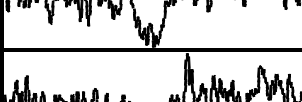
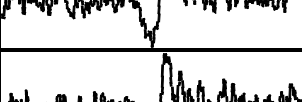


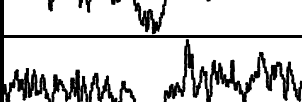
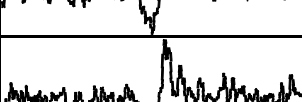
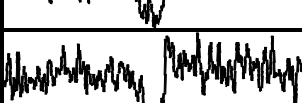
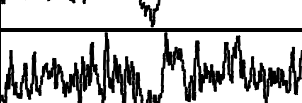


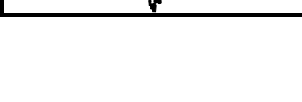
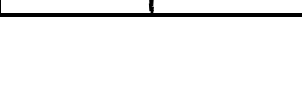
Organism	Number of genes	LCI(i,j) 5'	LCI(i,j) 3'	Index	M
<i>Mycoplasma leachii</i> pg50	882			2	75
<i>Mycoplasma cynos</i> c142	883			2	75
<i>Mycoplasma hyopneumoniae</i> 7448	657			2	73
<i>Mycoplasma hyopneumoniae</i> 168	693			2	73
<i>Mycoplasma hyopneumoniae</i> 168-1	683			2	73
<i>Mycoplasma hyopneumoniae</i> j	657			2	72
<i>Mycoplasma penetrans</i> hf-2	1037			2	71
<i>Mycoplasma synoviae</i> 53	659			2	71
<i>Mycoplasma putrefaciens</i> ks1	650			2	71
<i>Mycoplasma mobile</i> 163k	633			2	70
<i>Mycoplasma putrefaciens</i> mput9231	667			2	70
<i>Mycoplasma hominis</i> atcc 23114	523			2	69
<i>Mycoplasma pulmonis</i> uab ctip	782			2	68
<i>Mycoplasma conjunctivae</i> hrc581	692			2	67
<i>Mycoplasma genitalium</i> g37	475			2	62
<i>Mycoplasma arthritis</i> 158I3-1	631			2	62

Таблица 5. (Продолжение).

Organism	Number of genes	LCI(i,j) 5'	LCI(i,j) 3'	Index	M
<i>Mycoplasma gallisepticum</i> str r(low)	763			2	54
<i>Mycoplasma genitalium</i> m6282	484			2	53
<i>Mycoplasma gallisepticum</i> str r(high)	766			2	53
<i>Mycoplasma genitalium</i> m2288	506			2	52
<i>Mycoplasma genitalium</i> m6320	509			2	50
<i>Mycoplasma genitalium</i> m2321	499			2	50
<i>Mycoplasma gallisepticum</i> ny01 2001 047-5-1p	760			2	49
<i>Mycoplasma gallisepticum</i> ca06 2006 052-5-2p	763			2	49
<i>Mycoplasma gallisepticum</i> nc96 1596-4-2p	771			2	48
<i>Mycoplasma pneumoniae</i> fh	629			2	34
<i>Mycoplasma suis</i> ki3806	794			2	27
<i>Mycoplasma suis</i> str illinois	845			2	26
<i>Mycoplasma pneumoniae</i> m129-b7	613			2	25
Candidatus <i>Mycoplasma haemominutum</i> 'birmingham 1'	547			2	24
<i>Mycoplasma haemocanis</i> str illinois	1156			2	-6
<i>Mycoplasma haemofelis</i> ohio2	1527			3	6



Таблица 5. (Конец).

Organism	Number of genes	LCI(i,j) 5'	LCI(i,j) 3'	Index	M
<i>Mycoplasma haemofelis</i> str langford 1	1545			3	3
<i>Mycoplasma bovis</i> hubei-1	801			4	80
<i>Mycoplasma agalactiae</i> pg2	742			4	75
<i>Mycoplasma bovis</i> hb0801	814			4	69
<i>Mycoplasma bovis</i> pg45	765			4	67
<i>Mycoplasma agalactiae</i>	813			4	66
<i>Mycoplasma gallisepticum</i> str f	756			4	62
<i>Mycoplasma gallisepticum</i> nc08 2008 031-4-3p	739			4	53
<i>Mycoplasma gallisepticum</i> wi01 2001 043-13-2p	746			4	52
<i>Mycoplasma gallisepticum</i> nc06 2006 080-5-2p	744			4	51
<i>Mycoplasma gallisepticum</i> nc95 13295-2-2p	754			4	51
<i>Mycoplasma gallisepticum</i> va94 7994-1-7p	767			4	51
<i>Mycoplasma pneumoniae</i> 309	707			4	25
<i>Mycoplasma pneumoniae</i> ml29	688			4	25



**Таблица 6. Список проанализированных таксономических групп архей и виды, принадлежащие этим группам.**

**Thaumarchaeota**

*Cenarchaeum symbiosum* A uid61411

*Nitrosopumilus maritimus* SCM1 uid58903

**Euryarchaeota**

***Aciduliprofundum* (отдельный таксон)**

*Aciduliprofundum boonei* T469 uid433334

***Methanopyrales***

*Methanopyrus kandleri* AV19 uid57883

***Methanosarcinales***

*Methanococcoides burtonii* DSM 6242 uid58023

*Methanohalobium evestigatum* Z 7303 uid49857

*Methanohalophilus mahii* DSM 5219 uid47313

*Methanosaeta thermophila* PT uid58469

*Methanosarcina acetivorans* C2A uid57879

*Methanosarcina barkeri* Fusaro uid57715

*Methanosarcina mazei* Go1 uid57893

***Methanocellales***

*Methanocella arvoryzae* MRE50 uid61623

*Methanocella conradii* HZ254 uid157911

*Methanocella paludicola* SANAE uid42887

***Methanomicrobiales***

*Methanocorpusculum labreanum* Z uid58785

*Methanoculleus marisnigri* JR1 uid58561

*Methanoplanus petrolearius* DSM 11571 uid526954

*Methanoregula boonei* 6A8 uid588156

*Methanosphaerula palustris* E1 9c uid59193

*Methanospirillum hungatei* JF 1 uid58181

***Halobacteriales***

*Halalkalicoccus jeotgali* B3 uid50305

*Haloarcula marismortui* ATCC 43049 uid57719

*Halobacterium* NRC 1 uid57769

*Halobacterium salinarum* R1 uid615713

*Haloferax volcanii* DS2 uid46845

*Halomicrobium mukohataei* DSM 12286 uid59107

*Halopiger xanaduensis* SH 6 uid68105

*Haloquadratum walsbyi* C23 uid162019

*Haloquadratum walsbyi* DSM 16790 uid58673

*Halorhabdus utahensis* DSM 12940 uid59189

*Halorubrum lacusprofundi* ATCC 49239 uid58807

*Haloterrigena turkmenica* DSM 5511 uid43501

*Natrialba magadii* ATCC 43099 uid46245

*Natronomonas pharaonis* DSM 2160 uid58435

***Archaeoglobales***

*Archaeoglobus fulgidus* DSM 4304 uid57717

*Archaeoglobus profundus* DSM 5631 uid43493

*Archaeoglobus veneficus* SNP6 uid65269

*Ferroglobus placidus* DSM 10642 uid40863

***Methanobacteriales***

*Methanobacterium* AL 21 uid63623

*Methanobrevibacter ruminantium* M1 uid45857  
*Methanobrevibacter smithii* ATCC 35061 uid588273  
*Methanosphaera stadtmanae* DSM 3091 uid58407  
*Methanothermobacter marburgensis* Marburg uid51637  
*Methanothermobacter thermautotrophicus* Delta H uid57877

### **Thermococcales**

*Pyrococcus abyssi* GE5 uid62903  
*Pyrococcus furiosus* DSM 3638 uid57873  
*Pyrococcus horikoshii* OT3 uid57753  
*Pyrococcus yayanosii* CH1 uid68281  
*Thermococcus* 4557 uid708414  
*Thermococcus gammatolerans* EJ3 uid59389  
*Thermococcus kodakarensis* KOD1 uid58225  
*Thermococcus onnurineus* NA1 uid59043  
*Thermococcus sibiricus* MM 739 uid59399

### **Thermoplasmatales**

*Picrophilus torridus* DSM 9790 uid58041  
*Thermoplasma acidophilum* DSM 1728 uid61573  
*Thermoplasma volcanium* GSS1 uid57751

### **Methanococcales**

*Methanocaldococcus fervens* AG86 uid59347  
*Methanocaldococcus* FS406 22 uid42499  
*Methanocaldococcus infernus* ME uid48803  
*Methanocaldococcus jannaschii* DSM 2661 uid57713  
*Methanocaldococcus vulcanius* M7 uid41131  
*Methanococcus aeolicus* Nankai 3 uid588233  
*Methanococcus maripaludis* C5 uid587415  
*Methanococcus maripaludis* C6 uid589476  
*Methanococcus maripaludis* C7 uid588477  
*Methanococcus maripaludis* S2 uid58035  
*Methanococcus maripaludis* X1 uid70729  
*Methanococcus vanniellii* SB uid58767  
*Methanococcus voltae* A3 uid495293  
*Methanothermococcus okinawensis* IH1 uid51535

### **Crenarchaeota**

#### **Thermoproteales**

*Caldivirga maquilingensis* IC 167 uid58711  
*Pyrobaculum aerophilum* IM2 uid57727  
*Pyrobaculum arsenaticum* DSM 13514 uid58409  
*Pyrobaculum calidifontis* JCM 11548 uid58787  
*Pyrobaculum islandicum* DSM 4184 uid58635  
*Pyrobaculum neutrophilum* V24Sta uid58421  
*Pyrobaculum oguniense* TE7 uid84411  
*Thermofilum pendens* Hrk 5 uid58563  
*Vulcanisaeta distributa* DSM 14429 uid52827  
*Vulcanisaeta moutnovskia* 768 28 uid636317

#### **Sulfolobales**

*Metallosphaera cuprina* Ar 4 uid66329  
*Metallosphaera sedula* DSM 5348 uid587175  
*Sulfolobus acidocaldarius* DSM 639 uid58379  
*Sulfolobus acidocaldarius* DSM 639 uid58379  
*Sulfolobus islandicus* HVE10 4 uid162067

*Sulfolobus islandicus* LD85 uid43679  
*Sulfolobus islandicus* LS215 uid58871  
*Sulfolobus islandicus* M1425 uid58849  
*Sulfolobus islandicus* M1627 uid58851  
*Sulfolobus islandicus* M164 uid58841  
*Sulfolobus islandicus* YG5714 uid58923  
*Sulfolobus islandicus* YN1551 uid58825  
*Sulfolobus solfataricus* 982 uid1679989  
*Sulfolobus solfataricus* P2 uid57721

#### ***Desulfurococcales***

*Aeropyrum pernix* K1 uid57757  
*Desulfurococcus kamchatkensis* 1221n uid591331  
*Desulfurococcus mucosus* DSM 2162 uid622270  
*Hyperthermus butylicus* DSM 5456 uid57755  
*Ignicoccus hospitalis* KIN4 I uid583654  
*Ignisphaera aggregans* DSM 17230 uid51875  
*Pyrolobus fumarii* 1A uid734151  
*Staphylothermus hellenicus* DSM 12710 uid45893  
*Staphylothermus marinus* F1 uid58719  
*Thermosphaera aggregans* DSM 11486 uid48993

#### ***Acidilobales***

*Acidilobus saccharovorans* 345 15 uid513953

#### **Korarchaeota**

*Candidatus Korarchaeum cryptofilum* OPF8 uid58601

#### **Nanoarchaeota**

*Nanoarchaeum equitans* Kin4 M uid58009

**Таблица 7. Среднее значение оптимальной температуры среды обитания ( $T_{\text{опт}}$ ) 135 видов архей.**

Источники данных: БД по Methanogens (<http://metanogen.biotech.uni.wroc.pl/>), сайт <https://microbewiki.kenyon.edu/index.php/MicrobeWiki> и статьи (Amo et al., 2002; Antunes et al., 2008; Arab et al., 2000; Birrien et al., 2011; Blöchl et al., 1997; Bonch-Osmolovskaya et al., 1990; Borges et al., 2010; Burns et al., 2007; Burns et al., 2010; Cadillo-Quiroz et al., 2014; Castillo et al., 2007; Chan et al., 2013; De Rosa et al., 1975; Dopson et al., 2004; Elkins et al., 2008; Erauso et al., 1993; Fan et al., 2004; Fiala et al., 1986; Fiala and Stetter 1986; González et al., 1998; Gruber et al., 2004; Gumerov et al., 2011; Gutiérrez et al., 2007; Hafenbradl et al., 1996; Huber et al., 1989; Huber et al., 1997; Huber et al., 1998; Huber et al., 2006; Itoh et al., 1999; Itoh et al., 2002; Itoh et al., 2003; Jahn et al., 2004; Jahn et al., 2007; Jolivet et al., 2003; Klenk et al., 1997; Kublanov et al., 2009; Liu et al., 2011; Mardanov et al., 2012; McGenity et al., 1998; Miroschnichenko et al., 2001; Montalvo-Rodriguez et al., 1998; Nakagawa et al., 2004; Neuner et al., 1990; Niederberger et al., 2006; Oren et al., 1990; Oren et al., 2002; Perevalova et al., 2005; Perevalova et al., 2010; Preston et al., 1996; Prokofeva et al., 2009; Roh et al., 2007; Sako et al., 1996; Sako et al., 2001; Schleper et al., 1996; Schouten et al., 2008; Seob et al., 2006; Steinsbu et al., 2010; Suzuki et al., 2002; Torreblanca et al., 1986; Völkl et al., 1993; von Jan et al., 2010; Wainø et al., 2000; Yasuda et al., 1995; You et al., 2011; Zillig et al., 1991).

Организм	$T_{\text{опт}}$ среды обитания, °C
<i>Archaeoglobus veneficus</i> snp6	77,5
<i>Halalkalicoccus jeotgali</i> b3	35,5
<i>Haloarcula hispanica</i> atcc 33960	42,5
<i>Yaloarcula hispanica</i> n601	42,5
<i>Haloarcula marismortui</i> atcc 43049	45
<i>Halogeometricum borinquense</i> dsm 11551	40
<i>Haloquadratum walsbyi</i> c23	45
<i>Haloquadratum walsbyi</i> dsm 16790	45
<i>Halorhabdus tiamatea</i> sarl4b	45
<i>Halorhabdus utahensis</i> dsm 12940	50
<i>Metallosphaera cuprina</i> ar-4	65
<i>Methanocella arvoryzae</i> mre50	45
<i>Methanococcus maripaludis</i> c6	85
<i>Methanoregula boonei</i> 6a8	36
<i>Methanoregula formicica</i> smsp	31,5
<i>Pyrococcus yayanosii</i> ch1	98
<i>Thermococcus gammatolerans</i> ej3	88
<i>Thermococcus kodakarensis</i> kod1	85
<i>Thermococcus litoralis</i> dsm 5473	86,5
<i>Thermococcus onnurineus</i> na1	80
<i>Vulcanisaeta distributa</i> dsm 14429	87,5
<i>Acidilobus saccharovorans</i> 345-15	82,5
<i>Aciduliprofundum boonei</i> t469	67,5
<i>Caldisphaera lagunensis</i> dsm 15908	72,5
<i>Desulfurococcus fermentans</i> dsm 16532	81
<i>Desulfurococcus kamchatkensis</i> 1221n	85
<i>Desulfurococcus mucosus</i> dsm 2162	85
<i>Ferroplasma acidarmanus</i> fer1	38,5

<i>Fervidicoccus fontis</i> kam940	67,5
<i>Halobacterium salinarum</i> r1	37
<i>Halobacterium</i> sp. nrc-1	37
<i>Methanobacterium</i> sp al-21	34,5
<i>Methanocaldococcus vulcanius</i> m7	80
<i>Methanoculleus bourgensis</i> ms2	37
<i>Methanoculleus marisnigri</i> jr1	22,5
<i>Methanosalsum zhilinae</i> dsm 4017	45
<i>Methanosarcina barkeri</i> str. fusaro	35
<i>Methanosphaera stadtmanae</i> dsm 3091	38
<i>Methanosphaerula palustris</i> e1-9c	29
<i>Methanothermus fervidus</i> dsm 2088	83
<i>Nanoarchaeum equitans</i> kin4-m	90
<i>Picrophilus torridus</i> dsm 9790	60
<i>Pyrobaculum aerophilum</i> str. im2	100
<i>Pyrobaculum arsenaticum</i> dsm 13514	92,5
<i>Pyrobaculum calidifontis</i> jcm 11548	92,5
<i>Pyrobaculum oguniense</i> te7	92
<i>Pyrobaculum</i> sp. 1860	92,5
<i>Staphylothermus hellenicus</i> dsm 12710	85
<i>Staphylothermus marinus</i> fl	88,5
<i>Sulfolobus islandicus</i> l s 2 15	80
<i>Sulfolobus islandicus</i> m 16 27	80
<i>Thermogladius cellulolyticus</i> 1633	84
<i>Thermoplasma acidophilum</i> dsm 1728	57,5
<i>Thermoproteus uzoniensis</i> 768-20	90
<i>Acidianus hospitalis</i> w1	80
<i>Aeropyrum camini</i> sy1 = jcm 12091	85
<i>Aeropyrum pernix</i> k1	92,5
<i>Archaeoglobus fulgidus</i> dsm 4304	83
<i>Archaeoglobus profundus</i> dsm 5631	82
<i>Archaeoglobus sulfaticallidus</i> pm70-1	75
<i>Caldivirga maquilingensis</i> ic-167	85
<i>Candidatus korarchaeum cryptofilum</i> opf8	85
<i>Cenarchaeum symbiosum</i> a	10
<i>Ferroglobus placidus</i> dsm 10642	80
<i>Halomicrobium mukohataei</i> dsm 12286	42,5
<i>Halopiger xanaduensis</i> sh-6	37
<i>Halorubrum lacusprofundi</i> atcc 49239	38,5
<i>Halovivax ruber</i> xh-70	37
<i>Hyperthermus butylicus</i> dsm 5456	101
<i>Ignicoccus hospitalis</i> kin4i	90
<i>Ignisphaera aggregans</i> dsm 17230	93,5
<i>Metallosphaera sedula</i> dsm 5348	65

<i>Methanobacterium</i> sp. swan-1	34,5
<i>Methanobrevibacter ruminantium</i> m1	33
<i>Methanobrevibacter smithii</i> atcc 35061	38
<i>Methanobrevibacter</i> sp. abm4	37
<i>Methanocaldococcus fervens</i> ag86	85
<i>Methanocaldococcus infernus</i> me	85
<i>Methanocaldococcus jannaschii</i> dsm 2661	85
<i>Methanocaldococcus</i> sp. fs406-22	82,5
<i>Methanocella conradii</i> hz254	52,5
<i>Methanocella paludicola</i> sanae	36
<i>Methanococcoides burtonii</i> dsm 6242	23,4
<i>Methanococcus aeolicus</i> nankai-3	46
<i>Methanococcus maripaludis</i> c5	85
<i>Methanococcus maripaludis</i> c7	85
<i>Methanococcus maripaludis</i> s2	85
<i>Methanococcus maripaludis</i> x1	85
<i>Methanococcus vannieli</i> sb	30
<i>Methanococcus voltae</i> a3	38
<i>Methanocorpusculum labreanum</i> z	37
<i>Methanohalophilus mahii</i> dsm 5219	36
<i>Methanolobus psychrophilus</i> r15	18
<i>Methanomethylovorans hollandica</i> dsm 15978	35,5
<i>Methanoplanus petrolearius</i> dsm 11571	37,5
<i>Methanopyrus kandleri</i> av19	98
<i>Methanosaeta concilii</i> gp6	37,5
<i>Methanosaeta harundinacea</i> 6ac	35,5
<i>Methanosaeta thermophila</i> pt	57,5
<i>Methanosarcina acetivorans</i> c2a	37,5
<i>Methanosarcina mazei</i> go1	41
<i>Methanosarcina mazei</i> tuc01	41
<i>Methanospirillum hungatei</i> jf-1	41
<i>Methanothermobacter marburgensis</i> str. marburg	65
<i>Methanothermobacter thermautotrophicus</i> str. delta h	67,5
<i>Methanothermococcus okinawensis</i> ih1	62,5
<i>Methanotorris igneus</i> kol 5	88
<i>Natrinema pellirubrum</i> dsm 15624	38,5
<i>Natronomonas moolapensis</i> 8 8 11	38,5
<i>Pyrobaculum islandicum</i> dsm 4184	92,5
<i>Pyrobaculum neutrophilum</i> v24sta	92,5
<i>Pyrococcus abyssi</i> ge5	96
<i>Pyrococcus furiosus</i> com1	100
<i>Pyrococcus furiosus</i> dsm 3638	100
<i>Pyrococcus horikoshii</i> ot3	98
<i>Pyrolobus fumarii</i> 1a	106

<i>Sulfolobus acidocaldarius</i> dsm 639	85
<i>Sulfolobus acidocaldarius</i> n8	85
<i>Sulfolobus acidocaldarius</i> ron12i	85
<i>Sulfolobus acidocaldarius</i> susaz	85
<i>Sulfolobus islandicus</i> hve104	80
<i>Sulfolobus islandicus</i> l d 8 5	80
<i>Sulfolobus islandicus</i> lal141	80
<i>Sulfolobus islandicus</i> m 14 25	80
<i>Sulfolobus islandicus</i> m 16 4	80
<i>Sulfolobus islandicus</i> rey15a	80
<i>Sulfolobus islandicus</i> y g 57 14	80
<i>Sulfolobus islandicus</i> y n 15 51	80
<i>Sulfolobus solfataricus</i> 982	80
<i>Sulfolobus solfataricus</i> p2	80
<i>Sulfolobus tokodaii</i> str 7	77,5
<i>Thermococcus barophilus</i> mp	80
<i>Thermococcus sibiricus</i> mm 739	78
<i>Thermosphaera aggregans</i> dsm 11486	85
<i>Vulcanisaeta moutnovskia</i> 768-28	79

**Таблица 7. Среднее значение оптимальной температуры среды обитания ( $T_{\text{опт}}$ ) 135 видов архей (Конец).**